Estimation robuste en population finie et infinie

Cyril Favre-Martinoz

ANNÉE 2015

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Mathématiques et applications

Ecole doctorale Matisse

présentée par

Cyril FAVRE-MARTINOZ

préparée à l'Institut de Recherche Mathématique de Rennes
(IRMAR, UMR 6625)
et au Centre de Recherche en Economie et STatistique (CREST)

Estimation robuste en population finie et infinie

Thèse soutenue à l'Ensai
le 13 octobre 2015

devant le jury composé de :

Raymond CHAMBERS
Professeur, Université de Wollongong / rapporteur

François COQUET
Professeur, Ensai / directeur de thèse

Bernard DELYON
Professeur, Université Rennes 1 / examinateur

David HAZIZA
Professeur, Université de Montréal / co-directeur de thèse

Camélia GOGA
Maître de conférence HDR, Université de Bourgogne / rapporteur

Olivier SAUTORY
Inspecteur général de l'Insee / examinateur

Remerciements

Je tenais tout d'abord à remercier tout le personnel de l'Ensai pour leur accueil au sein de l'école, il y a maintenant près de 4 ans. Je pense tout particulièrement à Gaëlle Quéré avec qui j'ai collaboré lors de l'organisation du Colloque Francophone sur les Sondages et des différentes journées MAASC. Je remercie tous mes anciens camarades rennais Aline, Constance, David, Grégoire, Hugo, Madeline, Mathieu, Maxime, Nathan, Nicolas, Noémie, Pierre, Solène, Renaud, Vincent qui m'ont rappelé lors de leurs différents passages à Rennes qu'il y a une vie à côté de la recherche, je vous remercie pour tous les bons moments passés ensemble. Je n'oublie pas mes acolytes du Master de Statistique de Rennes 1, Alice, Morgane, Nicolas et Perrine ainsi que mes collègues doctorants rennais Emeline, François, Julie, Leslie, Quentin, Roman, Vincent et les autres doctorants non rennais, Anne, Aurélien, Hélène, Melike, Mélisande, Pierre-Olivier. J'ai une pensée tout particulière pour Maxime avec qui j'ai partagé des débats animés sur la Théorie des Sondages mais aussi sur la politique. Je n'oublie pas les enseignants en Statistique de l'Ensai Jann, Laura, Lionel, Magalie, Marian, Myriam, Nicolas, Pierre, Salima, Valentin, les enseignants en économie Nicolas, Samuel, Stéphane, Vincenzo et l'équipe du département informatique Alain, Jean-François, Laurence, Samuel, Thierry, Yann. Durant ces trois années de thèses, j'ai eu l'occasion de rencontrer des chercheurs très compétents, aux qualités relationnelles remarquables que je n'oublierai pas, je pense à Anne, Brigitte, Camélia, Daniel, Jean-Claude, Jean François, Jean Didier, Jocelyn, Hervé, Lionel, Nikos, Yves. Je terminerai par Eric et Guillaume, mes deux compères du laboratoire de statistique d'enquête avec qui j'ai passé quatre années très enrichissantes d'un point de vue personnel et professionnel, je vous remercie pour vos conseils avisés et le soutien que vous m'avez apporté au cours de cette thèse. Je remercie mon Directeur de thèse, François Coquet pour toutes les démarches effectuées afin de rendre cette thèse possible. Je lui suis reconnaissant pour le suivi régulier de mes travaux et pour toutes les connaissances qu'il m'a apportées sur la Statistique. Je remercie David Haziza, mon deuxième Directeur de thèse, pour les samedis ou les dimanches passés sur skype à discuter de résultats de simulation ou de tout autre chose, même à 6000 km tu as toujours été présent et je t'en suis grandement reconnaissant. Toujours rassurant, toujours enthousiaste, tu m'as fait comprendre que la recherche ne se limitait pas à résoudre des équations ou publier des articles. C'est une surtout une

passion qui se partage et se vie au quotidien. Je n'oublie pas mes deux collègues de bureau, Samuel et Gaspar, avec qui j'ai passé des heures à réinventer la statistique et partager tous les tracas du quotidien. Je remercie mon employeur l'Insee, qui m'a offert l'opportunité de faire une thèse et ses responsables qui m'ont soutenu dans ce projet, je pense à Pascal Chevalier, Laurent Di Carlo, Renan Duthion, Nicole Thomas, et plus particulièrement à Olivier Sautory qui a veillé à ce que ce projet arrive à terme. Je tenais à remercier mes parents, Brigitte et Pascal, pour leur soutien inconditionnel, leur patience, leurs précieux conseils et leur présence malgré les 850 km qui me séparait du nid familial. Je n'oublie pas ma bande d'acolyte Hugo, Jean-Mi, Nico, Pierre et Stéphane qui partagent ma vie depuis plus de 20 ans et qui m'ont rappelé qu'à côté de la thèse, il y a de la place pour un peu de vacances, du ski, des fous rires,et surtout de très belles amitiés. Enfin, comment ne pas parler de Mélody, ma compagne, je ne te remercierai jamais assez pour ta patience, malgré la distance, tu as toujours été présente et bienveillante.

Contents

List of Figures

List of Tables

Chapitre 1

Introduction

Un institut public de sondages comme l'Insee a pour objectif d'éclairer le débat public en collectant, produisant, analysant et diffusant des informations sur l'économie et la société française. A partir de données d'enquêtes recueillies sur des échantillons probabilistes judicieusement choisis, l'Insee souhaite estimer des grandeurs descriptives de l'économie ou de la population française. On peut s'intéresser par exemple à la distribution du patrimoine au sein des ménages, ainsi qu'aux taux de détention des différents actifs patrimoniaux, qui sont mesurés grâce à l'enquête patrimoine, ou encore à la part des investissements dans le chiffre d'affaire des PME françaises qui peut être évaluée par les enquêtes sectorielles annuelles (ESA). En Théorie des Sondages, ces grandeurs descriptives sont appelées des paramètres de population finie. Ils sont alors estimés à l'aide de différentes approches et différents estimateurs construits à partir des valeurs de la variable d'intérêt collectées sur l'échantillon et de l'information auxiliaire. En l'absence d'erreurs non dues à l'échantillonnage, les estimateurs classiquement utilisés en pratique sont sans biais, ou asymptotiquement sans biais, mais ils peuvent souffrir d'une variance très élevée lorsque la distribution de la variable d'intérêt est très asymétrique. C'est fréquemment le cas dans les enquêtes auprès des entreprises, lorsque l'on souhaite estimer des paramètres de population finie pour des variables économiques comme le chiffre d'affaire ou les montants d'investissements. Plus particulièrement, si on veut estimer le montant total des investis-

sements effectués en 2015 dans le secteur de l'industrie automobile française, il est indispensable d'inclure dans notre échantillon les entreprises qui contribuent à ce total de manière significative, comme Renault et Peugeot. En les sélectionnant d'office dans notre échantillon, on estime de façon certaine une grande partie du total des investissements, réduisant ainsi fortement la variance due à l'échantillonnage. Dans le cas contraire, ces unités ont un impact disproportionné sur l'estimation du paramètre considéré, et seront appelées dans la suite unités influentes. Ainsi, il est possible de se prémunir contre l'impact des valeurs influentes à l'étape du plan de sondage en sélectionnant d'office les unités potentiellement influentes. Dans les enquêtes auprès des entreprises, il est de coutume d'utiliser un plan stratifié aléatoire simple sans remise comportant une ou plusieurs strates exhaustives composées habituellement des grandes unités. Malheureusement, il est rarement possible d'éliminer complètement le problème des valeurs influentes à l'étape du plan de sondage. En effet, les strates dans les enquêtes auprès des entreprises sont habituellement formées au moyen d'une variable géographique, d'une variable de taille (par exemple, le nombre d'employés) et d'une variable de classification. Dans une enquête recueillant des dizaines de variables d'intérêt, il se peut que certaines d'entre elles soient peu ou pas liées aux variables de stratification, laissant apparaitre des valeurs influentes. C'est le cas notamment dans les enquêtes environnementales menées par Statistique Canada telle que l'enquête sur l'eau dans l'agriculture dont l'un des objectifs consiste à quantifier la quantité d'eau utilisée par les fermes canadiennes pour l'irrigation. Il s'avère que la consommation d'eau utilisée une année donnée est peu liée aux variables de stratification car la consommation est en partie expliquée par les conditions météorologiques locales subies par les fermes échantillonnées. Un autre problème conduisant à la présence de valeurs influentes dans l'échantillon est celui des migrants inter-strate plus connus sous le nom de "stratum jumpers" en anglais qui survient lorsque l'information auxiliaire disponible sur la base de sondage est différente de celle recueillie sur le terrain. Ces différences sont habituellement dues à des erreurs dans

la base de sondage (par exemple, dans le cas d'une base obsolète). Un stratum jumper est une unité qui n'appartient pas à la strate à laquelle elle aurait dû appartenir si l'information sur la base de sondage avait été correcte. Si une unité avec une grande valeur est assignée à une strate non-exhaustive, elle combinera alors une grande valeur de la variable d'intérêt et éventuellement un grand poids de sondage, ce qui la rendra potentiellement très influente. Pour traiter le problème de valeurs influentes, on a recours à des méthodes d'estimation robuste qui seront développées dans cette thèse et adaptées en fonction de la stratégie d'estimation choisie.

Dans ce mémoire, on aborde la question du traitement des valeurs influentes dans le but de produire des estimateurs robustes dans différents contextes d'estimation en Théorie des Sondages, mais aussi en population infinie.

Dans le chapitre 2, nous détaillons les notions essentielles de la Théorie des Sondages qui sont utilisées dans les chapitres suivants. Plus précisément, nous rappelons les notions de population finie, de paramètre d'intérêt, de plan de sondage et les différents approches qui coexistent en Théorie des Sondages. Nous passons en revue les différents estimateurs robustes existants dans la littérature, en détaillant auparavant les notions de valeurs aberrantes, valeurs influentes et valeurs extrêmes. Puis, nous définissons la notion de biais conditionnel, qui sera utilisé comme mesure d'influence et détaillons quelques propriétés de celui-ci. Enfin, nous ferons une brève revue de la littérature sur les méthodes utilisées dans le cas d'une estimation sur petits domaines.

La suite du manuscrit est rédigée sous forme d'articles presque indépendants, dans la mesure où chaque article traite d'un problème d'estimation robuste dans un contexte particulier en population finie ou infinie. Les méthodes d'estimation robuste utilisées dans les différents chapitres sont relativement similaires, car basées sur le biais conditionnel comme mesure d'influence. L'originalité apportée

dans chaque chapitre réside dans la définition et l'adaptation de cette notion de biais conditionnel suivant le contexte statistique adoptée.

Le chapitre 3 présente un travail réalisé en collaboration avec Jean-François Beaumont et David Haziza. Nous nous sommes intéressés aux méthodes dites de winsorisation, qui sont très souvent employées en Théorie des sondages pour traiter le problème des valeurs influentes et pour proposer des estimateurs robustes. Ces méthodes nécessitent la détermination d'une constante d'ajustement afin d'effectuer le compromis biais-variance. Dans la littérature, des choix optimaux de la constante d'ajustement ont été proposés pour un plan de sondage aléatoire simple sans remise stratifié par Kokic et Bell (1994) et étendus dans le cas d'un estimateur par le ratio par Clark (1995). Nous considérons une classe plus large d'estimateurs robustes qui contient entre autres les estimateurs winsorisés, et nous développons à partir du biais conditionnel des estimateurs robustes avec un choix adaptatif de la constante d'ajustement quelque soit le plan de sondage. Dans le contexte d'une estimation sur domaine, par exemple, lors de l'estimation du chiffre d'affaires des entreprises françaises par secteur d'activité, nous proposons une méthode permettant d'assurer la cohérence entre les estimations winsorisées calculées au niveau des domaines et l'estimateur winsorisé calculé au niveau de la population. Un article portant sur ce travail est publié dans la revue Techniques d'Enquêtes (2015).

Le chapitre 4 est également issu d'un travail réalisé en collaboration avec Jean-François Beaumont et David Haziza. L'objectif de ce chapitre est d'étendre les résultats proposés dans l'article de Beaumont et al. (2013) pour un plan de sondage à une seule phase à un plan de sondage à deux phases. Les plans de sondage à deux phases sont utilisés lorsqu'il y a peu ou pas d'information auxiliaire disponible à l'étape de construction du plan de sondage. Dans cette situation, il est souvent préférable de tirer un premier échantillon assez large en collectant de l'information peu coûteuse reliée à la variable d'intérêt que l'on souhaite mesurer.

A l'aide de cette information recueillie sur la première phase, on développe une stratégie de tirage efficace dans la deuxième phase, au cours de laquelle on tire un échantillon plus petit. Les plans en deux phases sont aussi très couramment utilisés pour modéliser la non-réponse : la première phase correspond à la phase d'échantillonnage classique et la deuxième phase correspond à la modélisation du mécanisme de non-réponse. Les estimateurs proposés dans le contexte d'estimation pour un plan de sondage à deux phases ou dans le cas de la modélisation de la non-réponse sont très sensibles aux valeurs influentes. C'est pourquoi nous proposons dans le chapitre 4 une stratégie d'estimation robuste dans ce contexte. Ce travail est en révision pour la revue Scandinavian Journal of Statistics (2014).

Au chapitre 5, nous détaillons un autre travail effectué en commun avec Jean-François Beaumont et David Haziza. L'estimation de la moyenne dans le cas d'une population asymétrique est un problème délicat en pratique. En effet, il est très courant d'observer des variables dont la distribution est asymétrique, c'est le cas par exemple du chiffre d'affaire des entreprises ou du revenu des ménages. En pratique, l'échantillon d'observation comprend souvent des unités qui sont très influentes sur la moyenne empirique, qui est l'estimateur souvent privilégié. Rivest (1994) propose un estimateur non paramétrique pour la moyenne d'une population asymétrique en winsorisant la plus grande ou les deux plus grandes observations de l'échantillon. Il montre que cet estimateur possède de bonnes propriétés en termes d'erreur quadratique moyenne. Sa stratégie consiste à réduire voire supprimer l'influence des plus grandes valeurs de l'échantillon. Notre démarche consiste à quantifier l'influence des unités de l'échantillon via le biais conditionnel adapté à un contexte de population infinie et de construire un estimateur robuste en réduisant l'impact des unités influentes identifiées. Nous donnons les propriétés de cet estimateur en termes d'erreur quadratique moyenne et nous développons une approximation de cette erreur quadratique moyenne suivant les différents domaines d'attraction possibles pour le maximum de la loi considérée.

Le chapitre 6 est issu d'un travail réalisé en collaboration avec Nikos Tzavidis et David Haziza sur l'estimation dans le cas du modèle linéaire généralisé ou du modèle linéaire mixte généralisé dans un contexte de population finie avec une extension au cas de l'estimation sur petits domaines. Les estimations reposants sur une approche modèle sont très sensibles à une mauvaise spécification du modèle ou à une mauvaise spécification de la distribution des erreurs. Dans le cas d'une variable d'intérêt continue, Sinha et Rao (2009) proposent une version robuste des estimateurs sur petits domaines en incorporant des estimations robustes des paramètres du modèle mixte. Chambers (2013) remarque que l'estimateur proposé par Sinha et Rao (2009) peut souffrir d'un biais important et propose une correction du biais à partir d'une approche similaire à Welsh et Ronchetti (1998). L'estimateur proposé par Chambers apporte une correction du biais pour les unités appartenant au domaine d'intérêt pour l'estimation, mais ne traite pas le biais engendré par les unités qui appartiennent à un autre petit domaine et qui ont une influence à travers l'estimation des coefficients et des effets aléatoires du modèle mixte. Une approche permettant une correction globale du biais a été proposée par Dongmo Jiongo et al. (2013), et s'appuie sur l'utilisation du biais conditionnel pour détecter puis traiter les valeurs influentes. Dans le cas de variables d'intérêt binaires ou discrètes, on a recours à des modèles logistiques mixtes ou des modèles de Poisson mixtes afin de proposer des estimations dans les petits domaines. On a recours à ces méthodes lorsque l'on souhaite, par exemple, estimer des taux de chômage pour de petites zones géographiques. Dans ce contexte, une approche bayésienne a été développée par Maiti (2001) et les estimateurs issus de la régression quantile développés par Chambers et Tzavidis (2006) ont été étendus au modèle binomial mixte et au modèle de Poisson mixte respectivement par Chambers, Salvati et Tzavidis (2014) et Tzavidis et al.(2014). Ces travaux seront détaillés dans le chapitre 6 et serviront de point de comparaison à nos estimateurs robustes construits à partir du biais conditionnel estimé dans un modèle de type

GLMM.

Chapitre 2

Rappel sur la Théorie des Sondages

Dans ce premier chapitre, nous ferons un bref rappel sur les bases de la Théorie des Sondages, qui nous seront utiles pour le développement des chapitres ultérieurs. Une grande partie de ces rappels s'appuient sur les ouvrages de Tillé (2001) et Ardilly (2006) pour les définitions usuelles utilisées en Théorie des Sondages, et sur les livres de Rao (2003) et Chambers (2012) pour l'estimation sur petits domaines.

2.1 Population

On se place dans le cadre d'une population finie notée U d'individus (ménages, entreprises,...) ou unités statistiques. Les unités de la population sont dites identifiables si elles peuvent être désignées par un numéro d'ordre ou un label de sorte que l'on notera simplement $U = \{1, ..., j, ..., N\}$. La notion de population correspond à ce qu'on appelle en pratique la base de sondage, soit une liste exhaustive de toutes les unités de la population. Pour préciser la base de sondage, on a recours à deux définitions. La définition de la population en compréhension est une définition conceptuelle de la base de sondage, par exemple : la population correspond à tous les français âgés de plus de 18 ans. La définition par extension correspond en général à une liste ou un fichier informatique comprenant toutes les unités.

En Théorie des Sondages, on s'intéresse à une variable d'intérêt (éventuellement vectorielle) y qui prend la valeur y_j sur l'individu j de la population U. Il est

important de remarquer que la variable y peut être déterministe ou être considérée comme une réalisation d'une variable aléatoire Y dont la loi est spécifiée dans un modèle de superpopulation.

2.2 Paramètre d'intérêt

En Théorie des Sondages, on cherche généralement à estimer un paramètre d'intérêt $\theta = \theta(y_j, j \in U)$, appelé aussi paramètre de population finie, qui est une fonction de $\boldsymbol{y}_N = (y_1, ..., y_j, ..., y_N)^\top$. Cette fonction est souvent linéaire en les valeurs prises par la variable d'intérêt sur la population : par exemple, si on souhaite estimer le total dans la population : $t_y = \sum_{j \in U} y_j$ ou si on souhaite estimer une moyenne : $\bar{y}_U = t_y/N$. Cependant, on peut déjà remarquer que l'estimation d'un total ou d'une moyenne sont deux problèmes d'estimation différents lorsque la taille de la population N est inconnue. On peut estimer $\bar{y}_U$ en estimant séparément le numérateur t_y et le dénominateur $N = \sum_{i \in U} 1$. On fait alors face à l'estimation d'un ratio, qui n'est plus linéaire. On peut également être amené à estimer d'autres paramètres tels que :

i) un taux ou un quotient de deux totaux, si l'on souhaite estimer par exemple le taux de chômage en France en 2015. On étudie alors le paramètre $\theta = t_y/t_x$ où t_y est le nombre total des chômeurs et t_x est le nombre total de personnes actives en France en 2015.

ii) un coefficient de corrélation entre deux variables d'intérêt x et y de la forme :

$$R = \frac{\sum_{j=1}^{N}(x_j - \bar{x}_U)(y_j - \bar{y}_U)}{\sqrt{\sum_{j=1}^{N}(x_j - \bar{x}_U)^2 \sum_{i=1}^{N}(y_j - \bar{y}_U)^2}}; \tag{2.2.1}$$

iii) un indice de Gini qui est un indicateur synthétique d'hétérogénéité pour une certaine variable d'intérêt y et dont les valeurs sont comprises entre 0 et 1. Il est égal à 0 dans une situation d'égalité parfaite où toutes les valeurs de la variable d'intérêt seraient égales. A l'autre extrême, il est égal à $1 - \frac{1}{N}$ dans une situation la plus inégalitaire possible, celle où toutes les valeurs de la variable sauf

une seraient nulles. Il est défini de la façon suivante :

$$G = \frac{2\sum_{j\in U} jy_j}{N\sum_{j\in U} y_j} - \frac{N+1}{N},$$

en supposant que les valeurs de y sont rangées par ordre croissant ($y_j \leq y_{j+1}$).

L'estimation des paramètres $i) - ii) - iii)$ est plus complexe que l'estimation d'un total dans la mesure où ils ne sont pas des fonctions linéaires de la variable d'intérêt. Dans ce cas, on a généralement recours à des méthodes de linéarisation introduites par Woodruff (1971) et développées par Binder (1983, 1996), Binder et Patak (1994), Wolter (1985), Deville (1999), Demnati et Rao (2004), Goga et al. (2009).

2.3 Plan de sondage

Le plan de sondage est une notion centrale en Théorie des Sondages. Il permet de définir comment et avec quelle probabilité les échantillons sont tirés.

Définition 2.1. *Un plan de sondage (non-ordonné) sans remise $p(.)$ est une loi de probabilité sur l'ensemble des parties de U, notée Ω.*

2.4 Quelques notions de statistique classique

On va redéfinir, dans le cadre de la Théorie des Sondages, les notions statistiques utilisées dans le cadre de la statistique classique. L'indice p indique que la notion est prise sous l'aléa du plan de sondage. L'indice m indique que la notion statistique est prise sous l'aléa défini dans le modèle.

Définition 2.2. *Soit $\hat{\theta}$ un estimateur de θ, on définit l'espérance par rapport au plan de sondage de l'estimateur $\hat{\theta}$ par :*

$$E_p\left(\hat{\theta}\right) = \sum_{s\in\Omega} \hat{\theta} p(s).$$

Le biais par rapport au plan de l'estimateur $\hat{\theta}$ pour le paramètre θ est défini par :

$$B_p(\hat{\theta}) = E_p(\hat{\theta}) - \theta.$$

Un estimateur $\hat{\theta}$ d'un paramètre d'intérêt θ est dit sans biais par rapport au plan si et seulement si

$$E_p(\hat{\theta}) = \theta \ \ pour \ \ tout \boldsymbol{y}_N \in \mathbb{R}^{\mathrm{N}}.$$

L'opérateur espérance permet de définir la variance par rapport au plan de l'estimateur $\hat{\theta}$:

$$V_p\left(\hat{\theta}\right) = E_p\left[\left\{\hat{\theta} - E_p\left(\hat{\theta}\right)\right\}^2\right].$$

Enfin, l'erreur quadratique moyenne par rapport au plan de l'estimateur $\hat{\theta}$ pour le paramètre θ est donnée par :

$$EQM_p(\hat{\theta}) = E_p\left\{(\hat{\theta} - \theta)^2\right\} = Var_p(\hat{\theta}) + B_p(\hat{\theta})^2.$$

2.5 Probabilité d'inclusion

On définit tout d'abord les variables indicatrices d'appartenance à l'échantillon qui prennent la valeur 1 si l'unité est dans l'échantillon et 0 sinon. Pour chaque individu, on définit :

$$I_j = \begin{cases} 1 & \text{si } j \in S \\ 0 & \text{si } j \notin S \end{cases}$$

Les I_j sont des variables aléatoires car ce sont des fonctions de la variable aléatoire S.

Définition 2.3. *A partir de ces variables aléatoires, on peut définir la probabilité d'inclusion, qui est la probabilité pour une unité d'appartenir à l'échantillon :*

$$\pi_j = E_p(I_j) = P(j \in S) = \sum_{s \ni j} p(s).$$

La probabilité d'inclusion d'ordre deux est la probabilité que deux unités distinctes apparaissent conjointement dans l'échantillon :

$$\pi_{jk} = E_p\left(I_j I_k\right) = P\left(j \in S \text{ et } k \in S\right) = \sum_{s \ni j,k} p\left(s\right) \text{ pour tout } j,k \in U,\ j \neq k.$$

Par convention, on notera $\pi_{jj} = \pi_j$. On définit enfin la matrice de variance-covariance des indicatrices I_j

$$\Delta_{jk} = \begin{cases} Cov_p\left(I_j, I_k\right) = E_p\left(I_j I_k\right) - E_p\left(I_j\right) E_p\left(I_k\right) = \pi_{jk} - \pi_j \pi_k & si\ j \neq k \\ Var_p(I_j) = E_p\left(I_j^2\right) - E_p\left(I_j\right)^2 = \pi_j\left(1 - \pi_j\right) & si\ j = k \end{cases} \tag{2.5.1}$$

Voici quelques exemples de plans de sondage classiques et les probabilités d'inclusion associées à ces plans.

2.5.1 Plan simple sans remise

Définition 2.4. *Un plan de taille fixe n est dit simple sans remise si et seulement si*

$$p(s) = \begin{cases} \binom{N}{n}^{-1} & si\ card(s) = n \\ 0 & sinon \end{cases}$$

Les probabilités d'inclusion de tous les ordres se déduisent du plan de sondage. Pour les probabilités d'inclusion d'ordre un, on a :

$$\pi_j = \sum_{s \ni j} p(s) = \sum_{s \ni j} \binom{N}{n}^{-1} = \binom{N-1}{n-1}\binom{N}{n}^{-1} = \frac{n}{N}, \text{ pour tout } j \in U.$$

Pour les probabilités d'inclusion d'ordre deux, on a :

$$\pi_{jk} = \sum_{s \ni j,k} p(s) = \sum_{s \ni j,k} \binom{N}{n}^{-1} = \binom{N-2}{n-2}\binom{N}{n}^{-1} = \frac{n(n-1)}{N(N-1)}, \text{ pour tout } j \neq k \in U. \tag{2.5.2}$$

Ainsi la matrice Δ_{jk} définit par l'expression (2.5.1) se réduit dans le cas d'un sondage aléatoire simple sans remise à :

$$\Delta_{jk} = \begin{cases} \pi_{jk} - \pi_j\pi_k = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n}{N}\left(1-\frac{n}{N}\right)\frac{1}{N-1} & \text{si } j \neq k \\ \pi_j(1-\pi_j) = \frac{n}{N}(1-\frac{n}{N}) = \frac{n}{N}\left(1-\frac{n}{N}\right) & \text{sinon.} \end{cases} \tag{2.5.3}$$

Il existe plusieurs procédures permettant la mise en œuvre du plan aléatoire simple sans remise ; voir Tillé (2001).

2.5.2 Plan simple stratifié sans remise

En pratique, on a plutôt recours à un sondage stratifié aléatoire simple sans remise afin d'améliorer la précision des estimations et de faciliter la mise en œuvre des enquêtes sur le terrain. On découpe la population U en H strates U_h de taille N_h telles que $U = \cup_{h=1}^{H} U_h$, $U_h \cap U_i = \emptyset, h \neq i$ et on effectue au sein de chacune de ces strates un sondage aléatoire simple sans remise de taille n_h de façon indépendante. On note S_h l'échantillon tiré dans la strate h.

Dans ce cas, on a

$$\pi_j = \frac{n_h}{N_h},\ j \in U_h$$

et

$$\Delta_{jk} = \begin{cases} \frac{n_h(N_h-n_h)}{N_h^2} & \text{si } j = k,\ j \in U_h \\ -\frac{n_h(N_h-n_h)}{N_h^2(N_h-1)} & \text{si } j \text{ et } k \in U_h,\ j \neq k \\ 0 & \text{si } j \in U_h,\ k \in U_i \text{ et } h \neq i \end{cases} \tag{2.5.4}$$

2.5.3 Plan de Poisson

Dans cette partie, nous allons décrire le plan de Poisson, qui est un plan de plus en plus utilisé dans les enquêtes auprès des entreprises.

Définition 2.5. *Dans un tirage de Poisson, chaque unité de la population U est sélectionnée de manière indépendante avec une probabilité π_j.*

Par conséquent, les variables aléatoires I_j, $j \in U$ sont indépendantes et elles suivent une loi de Bernoulli de paramètre π_j ; c'est-à-dire $I_j \sim \mathcal{B}(\pi_j)$. Il en découle

$$P(I_j = 1) = \pi_j,\ P(I_j = 0) = 1 - \pi_j.$$

La taille $n_s = \sum_{j \in U} I_j$ de l'échantillon est une variable aléatoire d'espérance $E_p(n_s) = \sum_{j \in U} \pi_j$ et de variance $Var_p(n_s) = \sum_{j \in U} \pi_j(1 - \pi_j)$.

La probabilité d'observer la réalisation s de la variable aléatoire S est :

$$p(s) = P(S = s) = \prod_{j \in s} \pi_j \prod_{j \in U - s} (1 - \pi_j).$$

Pour les probabilités d'inclusion d'ordre deux, on a :

$$\pi_{jk} = \pi_j \pi_k \ \ pour \ \ tout j \neq k \in U.$$

La matrice de variance-covariance des indicatrices I_j dans le cas d'un tirage poissonien se réduit à :

$$\Delta_{jk-Poiss} = \begin{cases} \pi_{jk} - \pi_j \pi_k = 0 & si\ j \neq k \\ \pi_j(1 - \pi_j) & sinon \end{cases}$$

L'implémentation du plan de Poisson consiste à générer N variables aléatoires indépendantes et identiquement distribuées selon une loi uniforme sur $[0, 1]$, notées $u_1, u_2, ..., u_N$. Si $u_j < \pi_j$ alors l'unité j est sélectionnée, sinon elle est rejetée et on passe à l'unité suivante.

L'intérêt du plan de Poisson est que son implémentation est extrêmement simple. Il est intéressant de noter que le plan de Poisson est celui qui maximise l'entropie.

Définition 2.6. *On appelle l'entropie d'un plan de sondage la quantité :*

$$I(p) = -\sum_{s \subset U} p(s) log(p(s)),$$

où on suppose que $0log(0) = 0$. *Plus l'entropie est élevée, plus l'ensemble des choix possibles pour la sélection de l'échantillon est large.*

2.6 Le π-estimateur

2.6.1 Estimation d'un total ou d'une moyenne

Définition 2.7. *L'estimateur de Horvitz et Thompson (1952) du total* t_y *est défini par*

$$\hat{t}_{y\pi} = \sum_{j \in S} \frac{y_j}{\pi_j}.$$

Cet estimateur est appelé le π-estimateur, estimateur de Horvitz-Thompson ou encore l'estimateur par dilatation. En effet, les valeurs prises par la variable d'intérêt y sur les unités échantillonnées sont dilatées par l'inverse des probabilités d'inclusion.

Remarque 2.1. Dans le cas d'un sondage stratifié aléatoire simple sans remise, on définit alors l'estimateur de Horvitz-Thompson par :

$$\hat{t}_{y\pi} = \sum_{h=1}^{H} \hat{t}_{y\pi_h},$$

où

$$\hat{t}_{y\pi_h} = \frac{N_h}{n_h} \sum_{j \in S_h} y_j.$$

Théorème 2.1. *Si* $\pi_j > 0$, *pour tout* $j \in U$, *alors* $\hat{t}_{y\pi}$ *estime* t_y *sans biais.*

Remarque 2.2. Si $\pi_j = 0$ pour au moins un individu j de la population, alors $\hat{t}_{y\pi}$

est biaisé par rapport au plan et son biais est donné par :

$$\begin{aligned} E_p(\hat{t}_{y\pi}) - t_y &= E_p\left(\sum_{j\in S}\frac{y_j}{\pi_j}\right) - t_y \\ &= E_p\left(\sum_{j\in U\,|\pi_j>0}\frac{I_j y_j}{\pi_j}\right) - t_y \\ &= \sum_{j\in U\,|\pi_j>0} y_j - t_y \\ &= t_y - \sum_{j\in U\,|\pi_j=0} y_j - t_y \\ &= \sum_{j\in U\,|\pi_j=0} y_j. \end{aligned} \tag{2.6.1}$$

Le biais (2.6.1) est appelé biais de couverture.

Remarque 2.3. Lorsque la taille N de la population est connue, un estimateur de la moyenne $\bar{y}_U$ dans la population est donné par $\hat{\bar{y}}_\pi = \hat{t}_{y\pi}/N$.

2.6.2 Variance du π-estimateur

La variance du π-estimateur est donnée par le théorème suivant :

Théorème 2.2. *Si $\pi_j > 0$ pour tout $j \in U$, alors*

$$Var_p(\hat{t}_{y\pi}) = \sum_{j\in U}\sum_{k\in U}\frac{y_j y_k}{\pi_j \pi_k}\Delta_{jk}. \tag{2.6.2}$$

Corollaire 2.1. *Dans le cas d'un sondage aléatoire simple sans remise, l'expression (2.6.2) se simplifie pour donner :*

$$Var_p(\hat{t}_{y\pi}) = N^2\frac{(1-\frac{n}{N})}{n}S^2_{y_U},$$

où $S^2_{y_U} = (N-1)^2\sum_{j\in U}(y_j - \bar{y}_U)^2$ désigne la dispersion de la variable y dans la population.

Théorème 2.3. *Si $\pi_{jk} > 0$ pour tout $(k, j) \in U \times U$, un estimateur sans biais de la variance du π-estimateur est donné par :*

$$\widehat{Var}(\hat{t}_{y\pi}) = \sum_{j\in S} \frac{y_j^2}{\pi_j^2}(1-\pi_j) + \sum_{j\in S}\sum_{k\in S,\, j\neq k} \frac{y_k y_j}{\pi_{jk}\pi_j\pi_k}\Delta_{jk}. \tag{2.6.3}$$

Dans le cas d'un sondage aléatoire simple sans remise, l'expression (2.6.3) se simplifie pour donner :

$$\widehat{Var}(\hat{t}_{y\pi}) = N^2 \frac{(1-\frac{n}{N})}{n} S_{y_S}^2, \tag{2.6.4}$$

où

$$S_{y_S}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y}_s)^2.$$

et

$$\bar{y}_s = \frac{1}{n}\sum_{k\in S} y_k.$$

Remarque 2.4. Dans le cas d'un sondage aléatoire simple, on a : $\hat{\bar{y}}_\pi = \bar{y}_s$.

Pour obtenir un estimateur de variance sans biais de l'estimateur de la moyenne, il suffit de remarquer que : $\widehat{Var}_p(\hat{\bar{y}}_\pi) = \widehat{Var}_p(\hat{t}_{y\pi})/N^2$.

2.7 Asymptotique en Théorie des Sondages

Comme en statistique inférentielle classique, il est intéressant de pouvoir évaluer la précision d'un estimateur. On estime souvent les variances des estimateurs afin d'en mesurer la qualité par le biais d'un intervalle de confiance. Ceci nécessite de connaitre la loi asymptotique de l'estimateur. Pour cela, on doit d'abord introduire ce que c'est que l'asymptotique dans le cadre de la Théorie des Sondages. Cette notion n'est pas aussi naturelle qu'en statistique inférentielle classique puisque la population est de taille finie. On doit avoir recours au modèle de superpopulation.

2.7.1 Le modèle de superpopulation

Le modèle de superpopulation de Isaki et Fuller (1982) introduit une population limite $U_{\mathbb{N}}$ avec un nombre infini dénombrable d'unités. La construction de la po-

pulation U_ν consiste en ν tirages indépendants. On tire une nouvelle réalisation de la variable d'intérêt y suivant le modèle de superpopulation spéficié pour la variable aléatoire Y pour obtenir la population $U_{\nu+1}$. On peut alors considérer une suite croissante de populations imbriquées telles que $U_1 \subset ... \subset U_{\nu-1} \subset U_\nu \subset U_{\nu+1} ... \subset U_{\mathbb{N}}$, de tailles $N_1 < N_2 < \cdots < N_\nu < ... < N_{\mathbb{N}}$, et une suite d'échantillons s_ν de taille n_ν, qui augmente avec ν, tirée dans U_ν selon le plan de sondage $p_\nu(s_\nu) = \mathbb{P}(s_\nu|U_\nu)$. On désigne par $\pi_{j(\nu)}$ et $\pi_{jk(\nu)}$ leur première et seconde probabilités d'inclusion. On peut remarquer que les populations sont croissantes et imbriquées, ce qui n'est pas nécessairement le cas de la suite des échantillons.

2.7.2 Convergence asymptotique

Les hypothèses généralement requises afin de développer les propriétés asymptotiques des estimateurs sont les suivantes :

$(H1)$ $\lim_{\nu\to+\infty} \frac{n_\nu}{N_\nu} = f_\nu = \pi \in]0,1[$;

$(H2) \forall v \in \mathbb{N}$, $\min_k \pi_{j(v)} \geq \lambda > 0$;

$\forall v \in \mathbb{N}$, $\min_{j\neq k} \pi_{kl(v)} \geq \lambda^* > 0$;

$\limsup_{\nu\to+\infty} n_\nu \max_{k\neq l} |\pi_{jk(v)} - \pi_{j(v)}\pi_{k(v)}| < C < +\infty$.

L'hypothèse (H1) suppose que la fraction de sondage tend vers une limite non dégénérée au fur et à mesure que la taille de la population U croît. Ainsi la taille de l'échantillon et la taille de la population augmentent à la même vitesse. L'hypothèse (H2) quantifie l'écart à l'indépendance des probabilités d'inclusion d'ordre 2 et impose aux $\pi_{k(v)}$ et $\pi_{kl(v)}$ d'être bornés inférieurement.

Définition 2.8. *Un estimateur $\hat{\theta}_\nu$ de θ_ν est dit consistant si pour tout $\epsilon > 0$*

$$\lim_{\nu\to+\infty} \mathbb{P}(|\hat{\theta} - \theta_\nu| > \epsilon|U_\nu) = 0.$$

2.7.3 Le théorème central limite

Le théorème central limite dans le cadre de la Théorie des Sondages n'a été démontré que pour certains plans de sondage. Erdös & Rényi (1959) et Hájek (1960) ont prouvé le résultat pour le sondage aléatoire simple sans remise.

Théorème 2.4. *Dans le cas du sondage aléatoire simple sans remise, en supposant que* $n_\nu \to +\infty$, $N_\nu - n_\nu \to +\infty$ *quand* $\nu \to +\infty$ *et que* S_ν^2 *désigne la dispersion dans la population* U_ν *on a :*

$$\sqrt{n_\nu}\frac{\hat{\mu}_\nu - \mu_\nu}{\sqrt{1 - f_\nu S_\nu^2}} \to \mathcal{N}(0,1) \text{ quand } \nu \to +\infty$$

si et seulement si la suite (Y_k) *vérifie la condition de Lindeberg-Hájek*

$$\lim_{N\to+\infty} \sum_{T_\nu(\delta)} \frac{Y_j - \mu_\nu}{(N_\nu - 1)S_\nu^2} = 0 \text{ quelque soit } \delta > 0,$$

où $T_\nu(\delta)$ *désigne l'ensemble des unités de* U_ν *pour lesquelles*

$$\frac{|Y_j - \mu_\nu|}{\sqrt{1 - f_\nu S_\nu^2}} > \delta\sqrt{n}.$$

La normalité asymptotique de l'estimateur de la moyenne dans le cas d'un sondage aléatoire simple stratifié a été discutée par Bickel et Freedman (1984) avec comme hypothèse que le nombre d'unités appartenant à chaque strate et le nombre d'unités échantillonnées tendent vers l'infini tout en supposant le nombre de strates fixé. La normalité asymptotique a aussi été démontrée par Krewski et Rao (1981) dans le cas d'un sondage aléatoire simple stratifié où le nombre de strates tend vers l'infini. Dans le cas des tirages à plusieurs degrés, Sen (1988) établit la normalité asymptotique d'un estimateur de type Horvitz-Thompson pour un échantillonnage successif des unités du premier degré. Pour une version plus détaillée de ces résultats, le lecteur pourra se référer à Thompson (1997). Plus récemment, le cas du sondage à deux phases a été traité dans l'article de Chen et Rao (2007).

2.8 Information auxiliaire

Dans la pratique, les instituts de statistique publics disposent de fichiers contenant de l'information collectée au cours d'enquêtes ou de recensements antérieurs. Elle est appelée information auxiliaire. Cette information peut être la connaissance de caractéristiques pour toutes les unités de la population ou elle peut être plus agrégée et être limitée à la connaissance de statistiques descriptives globales.

Dans la suite, on désigne par $\boldsymbol{X}$, la matrice de dimension $N \times p$ contenant l'ensemble de l'information auxiliaire disponible pour tous les individus de la population U et $\boldsymbol{x}_j = (x_1, .., x_p)^\top$ le vecteur des variables auxiliaires qui contient les p caractéristiques disponibles dans notre source d'information auxiliaire pour l'individu j, qui correspond à la j-ème ligne de la matrice $\boldsymbol{X}$. Ce supplément d'information peut permettre à la fois d'améliorer le plan de sondage initial s'il est disponible avant la mise en place de l'enquête mais il peut aussi être utilisé dans les procédures de redressement pour améliorer sous certaines conditions la précision des estimateurs classiques et vérifier certaines relations de cohérence.

2.9 Approche sous le plan, sous le modèle et approche assistée par le modèle

Dans le cadre de la Théorie des Sondages, il est important de distinguer trois approches inférentielles différentes : l'approche sous le plan, l'approche sous le modèle et l'inférence assistée par un modèle, que l'on retrouve sous les termes anglais "design-based approach", "model-based approach" et "model-assisted approach".

2.9.1 Approche sous le plan

Dans le cas d'une approche sous le plan, le paramètre d'intérêt que l'on souhaite estimer $\theta = \theta(y_j, j \in U)$ est un paramètre de population finie, en effet il ne dépend que des réalisations y_j, $j \in U$. Le vecteur des valeurs prises par la variable

d'intérêt $\boldsymbol{y_N}$ est considéré comme fixe. L'aléa réside uniquement dans le tirage de l'échantillon. L'espérance sous le plan peut être vue heuristiquement comme une moyenne sur l'ensemble des échantillons probabilistes possibles.

2.9.2 Approche sous le modèle

Dans le cas d'une approche sous le modèle, le vecteur des indicatrices $\boldsymbol{I}_N = (I_1, .., I_N)$ est fixé et les valeurs $y_1, .., y_j, ..., y_N$ de la variable d'intérêt observées sur la population U sont traitées comme des réalisations des variables aléatoires $Y_1, .., Y_j, ..., Y_N$ issues d'un modèle de superpopulation. On peut, par exemple, supposer que les réalisations $y_1, .., y_j, ..., y_N$ sont N réalisations indépendantes et identiquement distribuées de variables aléatoires d'espérance μ et de variance σ^2. Dans ce cas, μ et σ^2 sont des paramètres du modèle de superpopulation, ils sont inobservables et hypothétiques et ne peuvent être mesurés même si un recensement était effectué. Un modèle fréquemment utilisé dans le cas où l'information auxiliaire se limite à la connaissance d'un vecteur de variables explicatives $\boldsymbol{x}_j$ pour l'unité j s'écrit :

$$Y_j = m(\boldsymbol{x}_j, \boldsymbol{\beta}) + \epsilon_j, \tag{2.9.1}$$

où $\boldsymbol{\beta}$ représente le paramètre inconnu du modèle de superpopulation, m est une fonction appartenant à un espace de fonctions E de dimension finie et les ϵ_j sont des variables aléatoires indépendantes, identiquement distribuées d'espérance nulle et de variance fixée.

L'estimation convergente du paramètre $\boldsymbol{\beta}$ est traitée dans la littérature économétrique, voir par exemple Godambe et Thompson (1986) et Gary et al. (2014) . En Théorie des Sondages, on ne souhaite généralement pas décrire les liens au niveau du modèle de superpopulation, mais plutôt estimer les paramètres du modèle de superpopulation afin de caractériser des paramètres de population finie. Pour l'estimation des paramètres de population finie, on utilise les réalisations connues pour les unités sélectionnés dans l'échantillon tiré s. Les valeurs des unités non échantillonnées sont prédites à l'aide d'un modèle reliant les variables

aléatoires et l'information auxiliaire disponible $\boldsymbol{X}$. Dans le cas particulier d'un modèle linéaire, $m(\boldsymbol{x}_j, \boldsymbol{\beta}) = \boldsymbol{x}_j^\top \boldsymbol{\beta}$ et le modèle (2.9.1) se réduit à :

$$Y_j = \boldsymbol{x}_j^\top \boldsymbol{\beta} + \epsilon_j, \tag{2.9.2}$$

où

$$E\left(\epsilon_j | \boldsymbol{x}_j\right) = 0$$

$$Var(\epsilon_j | \boldsymbol{x}_j) = \sigma_j^2$$

$$\epsilon_j \text{ et } \epsilon_k \text{ indépendants lorsque } j \neq k$$

L'estimateur de type BLUP (Best Linear Unbiased Predictor) est alors défini (voir par exemple Valliant et al. (2000)) :

$$\hat{t}^{BLUP} = \sum_{j \in s} y_j + \sum_{j \in U \setminus s} \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}, \tag{2.9.3}$$

où $\hat{\boldsymbol{\beta}}$ est par exemple l'estimateur des moindres carrés du modèle de superpopulation, calculé sur les unités échantillonnées.

La notion d'informativité du plan de sondage est centrale dans le cas d'une approche sous le modèle. Elle permet d'obtenir des estimateurs consistants pour les paramètres de population finie sans prendre en compte les poids de sondage. Un plan de sondage est dit non-informatif ou ignorable pour l'estimation d'un paramètre si le modèle postulé sur la population est toujours valable au niveau de l'échantillon. Plus formellement (voir par exemple Skinner et al. (1989), Pfeffermann (1993), Valliant et al (2000)), on définit la fonction de densité conditionnelle de $\boldsymbol{Y}_s = (Y_1, .., Y_n)$ sachant l'échantillon $S = s$ par :

$$f_s(\boldsymbol{y}_s | \boldsymbol{x}_s; \theta) = f_U(\boldsymbol{y}_s | \boldsymbol{x}_s, S = s)$$

où $f_U(\boldsymbol{y}_s|\boldsymbol{x}_s;\theta)$ est la densité conditionnelle sur la population. Ce qui peut se réécrire avec la formule de Bayes, par :

$$f_s(\boldsymbol{y}_s|\boldsymbol{x}_s;\theta) = \frac{P(S=s|\boldsymbol{y}_s,\boldsymbol{x}_s) f_U(\boldsymbol{y}_s|\boldsymbol{x}_s;\theta)}{P(S=s|\boldsymbol{x}_s)}$$

Le plan $p(.)$ est dit non-informatif ou ignorable si :

$$f_s(\boldsymbol{y}_s|\boldsymbol{x}_s;\theta) = f_U(\boldsymbol{y}_s|\boldsymbol{x}_s;\theta) \tag{2.9.4}$$

Autrement dit, pour un paramètre de population finie θ donné, le plan $p(.)$ est dit non-informatif ou ignorable si la distribution conditionnelle de $\boldsymbol{Y}_s|\boldsymbol{x}_s$, $f_U(\boldsymbol{y}_s|\boldsymbol{x}_s;\theta)$, est égale à la distribution conditionnelle de $\boldsymbol{Y}_N = (Y_1,..,Y_N)^\top$ sachant $\boldsymbol{x}_N = (x_1,...,x_N)^\top$ restreinte aux unités échantillonnées, $f_U(\boldsymbol{y}_s|\boldsymbol{x}_s, S=s)$.

En particulier, lorsque

$$P(S=s|y_s,x_s) = P(S=s|x_s), \forall y_s, \tag{2.9.5}$$

l'égalité (2.9.4) est vérifiée. L'égalité des probabilités conditionnelles donnée dans l'expression (2.9.5) est vérifiée pour de nombreux plans. On peut citer, par exemple, le plan de sondage aléatoire simple sans remise et le plan de sondage équilibré tel que $\bar{x}_s = \bar{x}_U$.

2.9.3 Approche assistée par un modèle

Une dernière approche appelée approche assistée par un modèle est une approchée fréquemment mentionnée dans la littérature des Sondages ; l'idée sous-jacente à cette approche est d'utiliser un modèle de travail ("working model" en anglais) décrivant les liens entre la variable d'intérêt et l'information auxiliaire disponible dans le but d'optimiser le choix du plan de sondage et de l'estimateur. Cette approche est très bien décrite dans le livre de Särndal et al. (1992) et conduit dans le cas d'un modèle linéaire simple au célèbre estimateur GREG (General REGression estimator).

Si on utilise un modèle d'ajustement linéaire entre y et $\boldsymbol{X}$:

$$\forall j \in U,\ Y_j = \boldsymbol{x}_j^\top \boldsymbol{\beta} + e_k,$$

On peut alors décomposer le total t_y de la façon suivante :

$$t_y = \boldsymbol{t_x}^\top \boldsymbol{\beta} + t_e,$$

où $\boldsymbol{t}_x = \sum_{j \in U} \boldsymbol{x}_j$, $t_e = \sum_{j \in U} \left(y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta}\right)$ et $\boldsymbol{\beta}$ est défini par une méthode des moindres carrés pondérés par les poids c_j :

$$\boldsymbol{\beta} = \left(\sum_{j \in U} c_j^{-1} \boldsymbol{x}_j \boldsymbol{x}_j^\top\right)^{-1} \sum_{j \in U} c_j^{-1} \boldsymbol{x}_j y_j$$

l'estimateur GREG est défini par :

$$\hat{t}_y^{GREG} = \hat{t}_{y\pi} + \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}\right)^\top \hat{\boldsymbol{\beta}},$$

où

$$\hat{\boldsymbol{\beta}} = \left(\sum_{j \in s} c_j^{-1} d_j \boldsymbol{x}_j \boldsymbol{x}_j^\top\right)^{-1} \sum_{j \in s} c_j^{-1} d_j \boldsymbol{x}_j y_j.$$

Une version plus générale du GREG a été développée par Lehtonen et Veijanen (1998), Firth et Benett (1998) et Lehtonen et al. (2003, 2005) dans le cas de modèles assistés de type régression logistique ou régression logistique multinomiale et modèles mixtes. Enfin, Montanari et Ranalli (2002) propose une classe étendue d'estimateurs par la régression. Le traitement des valeurs influentes dans le cadre de cette approche n'est pas traité dans la thèse et constitue une perspective intéressante. Le cas particulier de l'estimateur GREG est traité dans l'article de Beaumont et al. (2013). De façon plus générale, on pourrait quantifier l'influence d'une unité en utilisant un biais conditionnel prenant à la fois en compte le modèle et le mécanisme d'échantillonnage, et proposé une version robuste en suivant une démarche proche de celle utilisée dans les différents chapitres de cette thèse. Une dernière difficulté serait de proposer des estimateurs robustes également calés,

comme par exemple Duchesne (1999), afin de proposer des estimateurs plus performants en présence de valeurs influentes et vérifiant les relations de cohérence classiques du calage.

Dans les chapitres 3 et 4, on se place dans le contexte d'une approche sous le plan, alors que dans le chapitre 6, on considère une approche modèle. Le chapitre 5 ne fait pas à appel à la notion de population finie, mais se place dans un contexte de statistique inférentielle classique.

2.10 Les méthodes d'estimation robuste

Nous nous attacherons à bien marquer la différence entre une valeur influente, une valeur aberrante et une valeur extrême. Pour bien comprendre ces différentes notions, nous allons donner quelques exemples dans le cas de la statistique classique, puis dans le cas de la Théorie des Sondages.

2.10.1 Valeurs aberrantes, valeurs extrêmes et valeurs influentes dans l'approche modèle en population infinie

Soit $X_1, X_2, ..., X_n$ un n échantillon de variables aléatoires indépendantes issu d'une loi $\mathcal{F}$. Soient $x_1, x_2, ..., x_n$ une réalisation de ces n variables aléatoires et $x_{(1)}, x_{(2)}, ..., x_{(n)}$ les données ordonnées dans l'ordre croissant.

Définition 2.9. Les valeurs $x_{(1)}$et $x_{(n)}$ sont respectivement l'observation extrême inférieure et supérieure de l'échantillon.

Autant, il est relativement aisé de définir la notion de valeurs extrêmes d'un échantillon, autant il est plus difficile de définir de façon mathématiquement précise les notions de valeurs aberrantes et valeurs influentes. Il y a deux raisons majeures. Le fait de définir une unité comme aberrante et/ou influente dépend du contexte, c'est-à-dire du problème que le statisticien souhaite modéliser, mais

aussi des méthodes statistiques utilisées. Ces notions sont aussi très subjectives, car elles sont surtout relatives à l'appréciation de l'utilisateur.

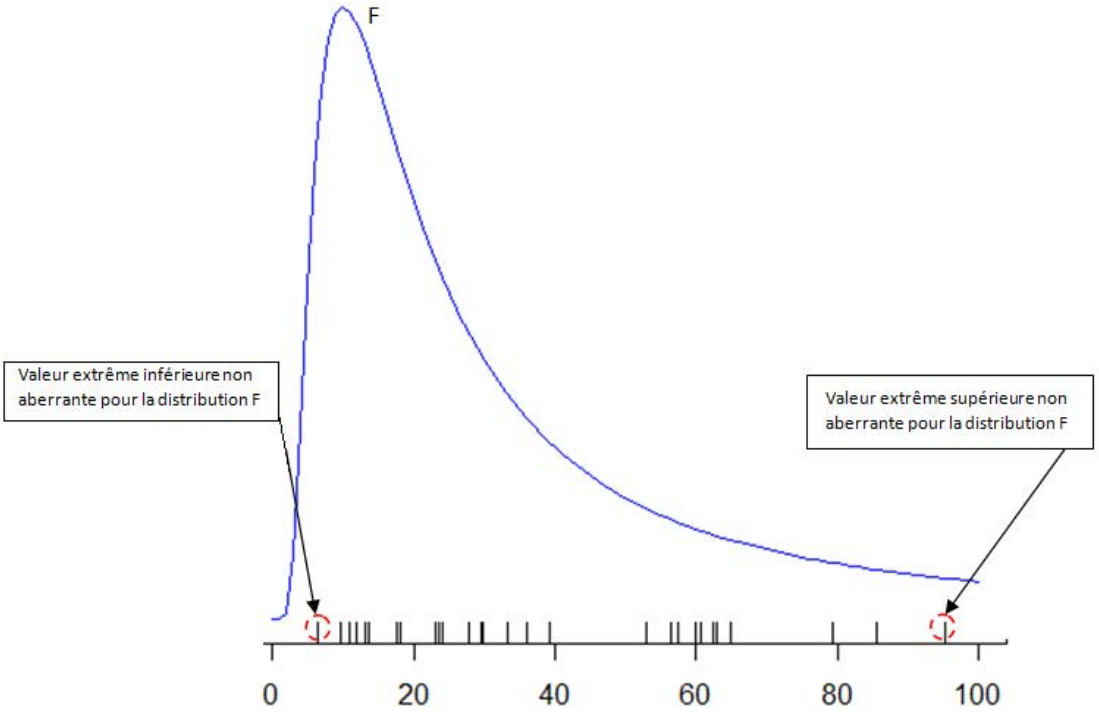
Une donnée sera considérée comme aberrante si elle n'est pas issue du même modèle que celui qui tient pour la majorité des données .

Considérons maintenant un paramètre d'intérêt θ et $\hat{\theta}$ un estimateur de θ, une valeur sera influente pour l'estimateur $\hat{\theta}$ du paramètre θ si elle a un impact significatif sur l'erreur quadratique moyenne de l'estimateur $\hat{\theta}$. Le terme "significatif" traduit la subjectivité présente dans cette définition

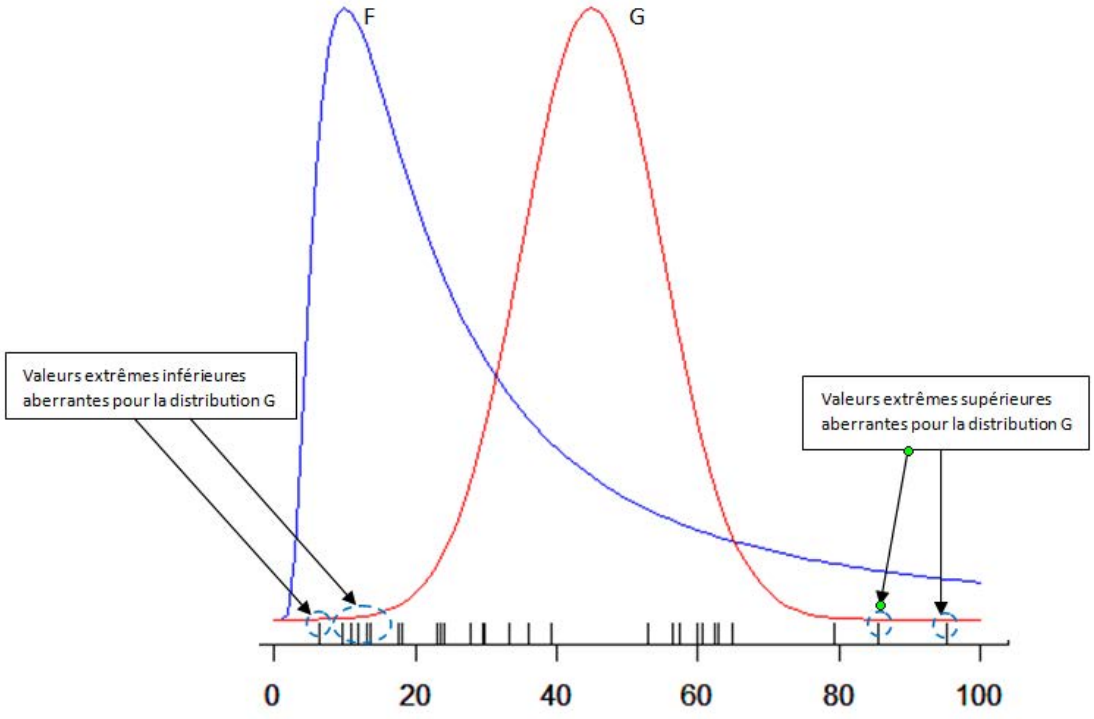
La notion de valeur aberrante est conditionnelle à la distribution ou au modèle considérés. Pour mieux comprendre cette idée, voici quelques exemples :

Dans la figure (2.10.1a), si on considère que notre échantillon est issu d'une loi de Fréchet F , dont la densité est représentée en bleu sur le graphique, alors les valeurs extrêmes de l'échantillon ne sont pas des valeurs aberrantes.

Sur la deuxième figure (2.10.1b), si on considère que notre échantillon est issu d'une loi de Normale G, dont la densité est représentée en rouge sur le graphique, alors les valeurs extrêmes de l'échantillon sont des valeurs aberrantes : la probabilité d'observer de telles valeurs sachant que notre échantillon est issu d'une loi Normale est extrêmement faible. Ainsi on voit que la notion de valeur aberrante est relative à une certaine distribution.



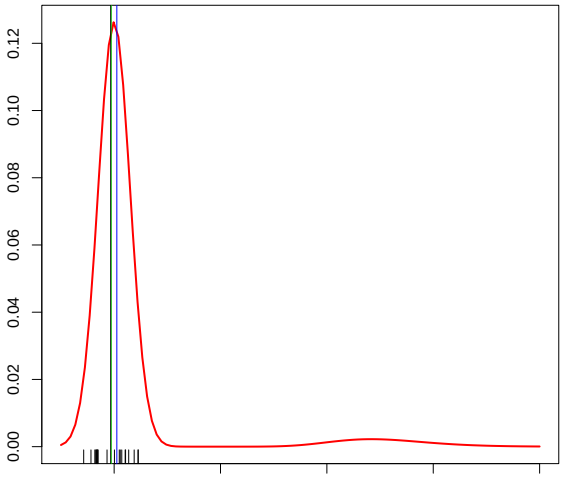
(a) Échantillon modélisé avec une loi normale



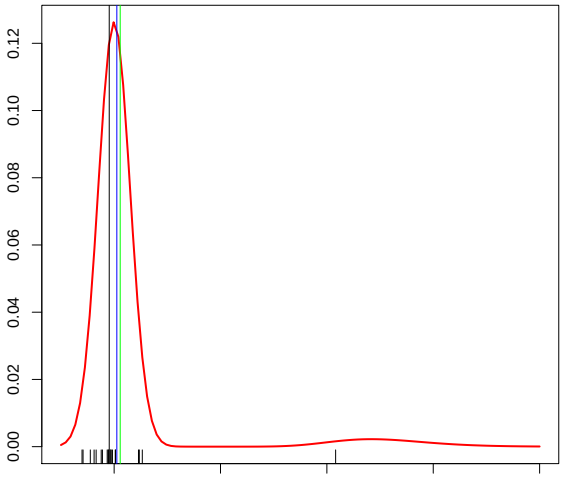
(b) Échantillon modélisé avec une loi normale ou une loi de Fréchet

FIGURE 2.10.1: Exemples de valeurs extrêmes

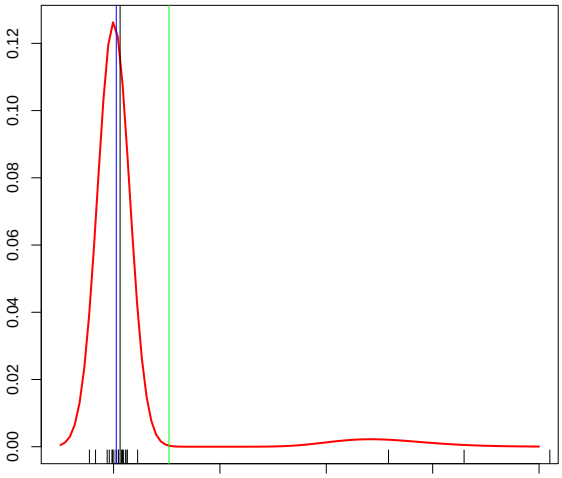
La notion de valeur influente est directement liée au paramètre que l'on souhaite estimer et à la façon de l'estimer. Supposons, par exemple que notre modèle soit un modèle de mélange du type $Y_i = (I_i - 1)N_i + I_i L_i$, où N_i est une variable aléatoire suivant une loi Normale de moyenne 0 et de variance 9, L_i est une variable aléatoire suivant une loi log normale de paramètre 50 et 1.2 (autrement dit $\log(L_i)$ suit une loi Normal de moyenne 50 et de variance 1.2), et I_i est une variable aléatoire de Bernoulli de paramètre $p = 0.05$. On souhaite estimer le paramètre $\mu_1 = E_F(Y_i)$. Dans notre exemple, $\mu_1 = 2.541899$. Les graphiques suivants présentent le positionnement de la moyenne empirique (en vert), de la médiane empirique (en noir) suivant différentes réalisations du modèle par rapport au vrai paramètre du modèle μ_1 représenté en noir. Pour la moyenne empirique, les valeurs extrêmes de l'échantillon ont un impact significatif sur la stabilité de l'estimateur, elles sont par conséquent considérées comme influentes. Si on considère la médiane empirique, les valeurs extrêmes n'ont aucun impact sur celle-ci, étant donné son point de rupture élevée, elles ne sont pas considérées comme influentes. La notion de valeurs influentes est liée à l'estimateur considéré et à ses propriétés.



(a) Réalisation échantillon N° 1



(b) Réalisation échantillon N° 2



(c) Réalisation échantillon N° 3

FIGURE 2.10.2: Exemples de valeurs influentes

2.10.2 Valeurs aberrantes, valeurs extrêmes et valeurs influentes en population finie

En statistique classique, on est en présence de populations infinies dont on cherche, par exemple, à estimer la moyenne, comme dans l'exemple précédent. Dans ce cas, une valeur aberrante est une valeur qui a été générée selon un modèle différent de celui qui a généré la majorité des observations. La présence de valeurs aberrantes dans l'échantillon peut s'expliquer par le fait que la population dont est générée l'échantillon est un mélange de distributions ou encore que certaines observations sont sujettes à des erreurs de mesure. En statistique classique, on cherche habituellement à mener des inférences sur la population des valeurs non aberrantes. Le but est donc de construire des estimateurs robustes au sens où ces derniers sont peu affectés par la présence de données aberrantes dans l'échantillon. Dans ce contexte, il est préférable de construire des estimateurs robustes ayant un point de rupture élevé et/ou une fonction d'influence bornée. En population finie, les erreurs de mesure sont corrigées à l'étape de la vérification si bien que l'on suppose qu'il n'en reste plus à l'étape de l'estimation. Le but est de mener une inférence sur la population "totale" qui comprend les valeurs aberrantes ainsi que les valeurs non aberrantes. Autrement dit, contrairement à la statistique classique, on ne s'intéresse pas qu'à la population de valeurs non aberrantes. Dans ce contexte, des estimateurs affichant un point de rupture élevé et/ou une fonction d'influence bornée ne sont généralement pas appropriés car ils peuvent conduire à des biais importants. C'est pourquoi, il est nécessaire de distinguer les notions de valeurs aberrantes et valeurs influentes suivant le contexte d'estimation dans lequel on se place. En Théorie des Sondages, une difficulté supplémentaire apparait, car il faut également tenir compte de l'approche utilisée pour estimer les paramètres d'intérêt.

Dans le cas *design-based* ou approche sous le plan, la notion de valeur aberrante n'a de sens que lorsqu'on observe une erreur de mesure. Un exemple typique d'une valeur aberrante dans le cas d'une enquête entreprise est la valeur du chiffre

d'affaire d'une entreprise indiquée en euro au lieu d'être indiquée en milliers d'euros.

Il est important de distinguer dans notre échantillon, deux types d'unités aberrantes : les unités aberrantes représentatives et les unités aberrantes non représentatives. Ce concept d'unités représentatives a été introduit et discuté par Chambers (1986) dans le cas d'une approche *model-based*. Les unités représentatives sont des unités dont la valeur collectée sur l'échantillon est correcte et n'est pas considérée comme unique au sens où il est probable qu'il existe dans notre population U d'autres unités ayant une valeur collectée du même ordre de grandeur. Dans le cas de l'estimation d'un paramètre de population finie comme un total, ces unités ont une importance considérable dans l'estimation de celui-ci et on ne peut pas se permettre de leur mettre un poids égal à 1, car cela reviendrait à les considérer comme uniques. Les valeurs aberrantes non représentatives sont des unités dont la valeur collectée est erronée, à cause d'un dysfonctionnement dans le processus de collecte : un cas classique est le chiffre d'affaire d'une entreprise indiqué en euro au lieu d'être indiqué en milliers d'euros. Le traitement de ce type d'unité peut être corrigée à l'étape d'apurement des données, notamment par des processus d'imputation. Ces unités aberrantes sont de par le fait uniques, et on peut leur attribuer un poids de 1 dans la suite du processus d'estimation ou corriger leur valeur si on est capable d'identifier l'erreur.

Avant de définir la notion de valeur influente de façon générale, quelque soit l'approche considérée, sous le modèle ou sous le plan, on introduit le concept de configuration :

Définition 2.10. *Dans le cas d'une approche sous le plan, une configuration $\mathcal{C}$ est définie par le quadruplet suivant :*
(1) une population U, ou une variable d'intérêt y ;
(2) un paramètre d'intérêt ;
(3) un plan de sondage ;

(4) un estimateur.

Définition 2.11. *Dans le cas d'une approche modèle, la notion de configuration est définie par le quadruplet suivant :*
(1) une population U, ou une variable d'intérêt y ;
(2) un modèle (m) ;
(3) un paramètre d'intérêt ;
(4) un prédicteur.

Dans une configuration $\mathcal{C}$ donnée, une valeur sera définie comme influente si elle a un impact significatif sur l'erreur quadratique moyenne de l'estimateur considéré.

Dans le chapitre 3, nous placerons dans la configuration d'une population quelconque, avec un paramètre d'intérêt de type moyenne ou total. Le plan de sondage sera lui-aussi quelconque et l'estimateur considérée sera l'estimateur Horvitz-Thompson. Dans le chapitre 4, nous nous placerons dans la configuration d'une population quelconque, avec un paramètre d'intérêt de type moyenne ou total, le plan de sondage considéré sera un plan de sondage à deux phases, avec dans le cas de la non-réponse une deuxième phase poissonienne. L'estimateur considéré sera l'estimateur par double dilatation ou l'estimateur ajusté par la méthode des scores en présence de non-réponse. Dans le chapitre 5, nous utiliserons également une approche modèle mais dans un contexte de statistique classique avec comme paramètre d'intérêt l'espérance de la distribution considérée et un estimateur de type moyenne empirique. Dans le chapitre 6, nous utiliserons une approche modèle, avec un modèle linéaire généralisé ou un modèle linéaire mixte généralisé, pour l'estimation d'un paramètre de type moyenne ou total avec un estimateur de type prédiction. Dans le cadre d'une approche sous le plan, les estimateurs classiques considérés (Horvitz-Thompson, estimateur par le ratio, estimateur par le calage) sont sans biais ou approximativement sans biais. Dans le cadre d'une approche modèle, l'estimateur classique considéré est l'estimateur BLUP. Dans

les deux approches, ces estimateurs sont très instables en présence de valeurs influentes. Il s'agit donc traiter à l'aide d'une certaine mesure d'influence ces unités influentes et de produire à partir de cette mesure des estimateurs plus stables.

2.10.3 Le biais conditionnel comme mesure d'influence

Le concept de configuration est important dans la mesure où une unité est influente dans une configuration donnée : c'est-à-dire qu'une unité est influente pour un plan, un paramètre et un estimateur ou pour un modèle (m), un paramètre et un estimateur. Il s'agit maintenant de se donner une mesure d'influence qui prend en compte le plan de sondage mais aussi le modèle que l'on impose à nos données.

Dans le cas d'une approche modèle la notion de biais conditionnel comme mesure d'influence a été introduit par Muñoz-Pichardo et al. (1995). Dans cet article, ils développent le biais conditionnel dans le cas d'un modèle linéaire. Dans le cas d'une approche sous le plan, la notion de biais conditionnel a été développé dans deux articles de Moreno-Rebollo et al. (1995, 1999). Cette notion de biais conditionnel a été reprise par Beaumont et al. (2013) afin de quantifier l'influence sous le plan ou sous le modèle d'une unité afin de construire ensuite des estimateurs robustes.

2.10.3.1 Biais conditionnel pour une approche sous le plan

Définition 2.12. [*Biais conditionnel pour une approche sous le plan*]

Soit $U = (1, ..., k, ..., N)$ une population finie, $p(S)$ un plan de sondage défini sur U et Y la variable d'intérêt à observer sur la population. Soit θ le paramètre d'intérêt et $\hat{\theta}$ un estimateur de θ .

Le biais conditionnel d'une unité échantillonnée i associé à l'estimateur $\hat{\theta}$ est défini par :

$$B_i^{\hat{\theta}}(I_i = 1) = \mathbb{E}_P\left(\hat{\theta}|I_i = 1\right) - \mathbb{E}_P\left(\hat{\theta}\right),$$

où I_i est la variable indicatrice d'appartenance à l'échantillon qui prend la valeur 1 si l'unité i est dans l'échantillon et 0 sinon. De façon similaire, on définit le

biais conditionnel d'une unité i non échantillonnée associé à l'estimateur $\hat{\theta}$ par :

$$B_i^{\hat{\theta}}(I_i = 0) = \mathbb{E}_P\left(\hat{\theta}|I_i = 0\right) - \mathbb{E}_P\left(\hat{\theta}\right).$$

Le biais conditionnel est une mesure d'une influence car il permet d'observer l'impact moyen engendré sur l'estimateur par le fait de contraindre l'unité i à appartenir ou non à l'échantillon.

Proposition 2.1.

$$B_i^{\hat{\theta}}(I_i = 0) = -\frac{\pi_i}{1-\pi_i} B_i^{\hat{\theta}}(I_i = 1).$$

Démonstration. Il suffit d'utiliser la propriété de l'espérance conditionnelle. On a $\mathbb{E}_P\left(\hat{\theta}\right) = \mathbb{E}_P\left\{\mathbb{E}_P\left(\hat{\theta}|I_i\right)\right\}$.

Or,

$$\mathbb{E}_P\left\{\mathbb{E}_P\left(\hat{\theta}|I_i\right)\right\} = \pi_i \mathbb{E}_P\left(\hat{\theta}|I_i = 1\right) + (1-\pi_i)\mathbb{E}_P\left(\hat{\theta}|I_i = 0\right).$$

On a alors :

$$\pi_i \mathbb{E}_P\left(\hat{\theta}\right) + (1-\pi_i)\mathbb{E}_P\left(\hat{\theta}\right) = \pi_i \mathbb{E}_P\left(\hat{\theta}|I_i = 1\right) + (1-\pi_i)\mathbb{E}_P\left(\hat{\theta}|I_i = 0\right).$$

Après réarrangement des termes, on obtient :

$$B_i^{\hat{\theta}}(I_i = 0) = -\frac{\pi_i}{1-\pi_i} B_i^{\hat{\theta}}(I_i = 1).$$

□

L'influence d'une unité sur un estimateur sera quantifiée à l'aide du biais conditionnel de cet unité. On va calculer explicitement cette influence sur l'estimateur de Horvitz-Thompson.

Proposition 2.2. *Le biais conditionnel d'une unité échantillonnée pour l'estimateur de Horvitz-Thompson est donné par :*

$$\begin{aligned} B_i^{HT}(I_i = 1) &= E_p\left(\hat{t}_{y\pi}|I_i = 1\right) - t_y \\ &= \sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right) y_j. \end{aligned} \tag{2.10.1}$$

Démonstration.

$$\begin{aligned} E_p(\hat{t}_{y\pi}|I_i=1) &= E_p\left(\sum_{j\in U}\frac{y_jI_j}{\pi_j}|I_i=1\right) \\ &= \sum_{j\in U}\frac{y_j}{\pi_j}E_p(I_j|I_i=1) \\ &= \sum_{j\in U}\frac{y_j}{\pi_j}\sum_{l=0}^{1} l\frac{P(I_j=l,I_i=1)}{P(I_i=1)} \\ &= \sum_{j\in U}\frac{y_j}{\pi_j}\frac{\pi_{ij}}{\pi_i} \\ &= \sum_{j\in U}\frac{\pi_{ij}}{\pi_i\pi_j}y_j. \end{aligned}$$

On a donc

$$B_i^{HT}(I_i=1) = \sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)y_j.$$

□

Nous allons maintenant calculer le biais conditionnel pour deux tirages classiques : le tirage poissonien et le tirage aléatoire simple stratifié sans remise.

Exemple 2.1. Pour un tirage poissonien, le biais conditionnel associé à l'estimateur de Horvitz-Thompson pour les unités échantillonnées est donné par :

$$B_i^{HT}(I_i=1) = (d_i-1)y_i,$$

où d_i désigne l'inverse de la probabilité d'inclusion associée à l'unité i, $d_i = \pi^{-1}$.

Dans le cas poissonien, on a $\pi_{ij} = \pi_i\pi_j$ si $i \neq j$ et $\pi_{ij} = \pi_i$ si $i = j$. L'expression (2.10.1) se simplifie pour donner

$$B_i^{HT}(I_i=1) = \left(\frac{1}{\pi_i}-1\right)y_i = (d_i-1)y_i. \tag{2.10.2}$$

Exemple 2.2. Pour un sondage aléatoire simple sans remise, le biais conditionnel associé à l'estimateur de Horvitz-Thompson pour les unités échantillonnées est donné par :

$$B_i^{HT}(I_i = 1) = \frac{N}{N-1}\left(\frac{N}{n} - 1\right)(y_i - \bar{y}_U). \tag{2.10.3}$$

Ce résultat se démontre en remplaçant les probabilités d'inclusion générales dans l'expression (2.10.1) par les probabilités d'inclusion du plan aléatoire simple sans remise données par l'égalité (2.5.2) .

Exemple 2.3. Pour un tirage aléatoire simple stratifié sans remise, le biais conditionnel associé à l'estimateur de Horvitz-Thompson pour les unités échantillonnées dans la strate h est donné par :

$$B_i^{HT}(I_i = 1) = \frac{N_h}{N_h - 1}\left(\frac{N_h}{n_h} - 1\right)(y_{hi} - \bar{y}_{U_h})$$

où N_h est la taille de la population U_h de la strate h, n_h est la taille de l'échantillon S_h tiré dans la strate h et $\bar{y}_{U_h}$ est la moyenne dans la strate h : $\bar{y}_{U_h} = N_h^{-1}\sum_{j\in U_h} y_{hj}$.

Exemple 2.4. Pour les plans à grande entropie, une approximation du biais conditionnel associé à l'estimateur de Horvitz-Thompson est fournie dans l'article de Beaumont et al. (2013) :

$$B_i^{HT}(I_i = 1) \simeq (d_i - 1)\left[\left\{1 + D^{-1}\pi_i(1-\pi_i)\right\} y_i - \phi\pi_i\right],$$

où $D = \sum_{l\in U}\pi_l(1-\pi_l)$ et $\phi = D^{-1}\sum_{j\in U}(1-\pi_j)\,y_j$.

Proposition 2.3. *Quel que soit le plan de sondage $p(.)$, on a la relation suivante :*

$$Var_p(\hat{t}_{y\pi}) = \sum_{j\in U}\sum_{k\in U}\frac{y_j y_k}{\pi_j\pi_k}\Delta_{kl} = \sum_{i\in U} B_i^{HT}(I_i = 1)y_i. \tag{2.10.4}$$

Démonstration. Immédiate en utilisant l'expression (2.10.1) du biais conditionnel. □

La variance de l'estimateur de Horvitz-Thompson est donc directement reliée au biais conditionnel et on constate qu'une unité ayant un fort biais conditionnel contribuera fortement à la variance. De plus, elle contribuera d'autant plus fort à la variance que la valeur de la variable d'intérêt y_i sera élevée.

Proposition 2.4. *L'erreur d'échantillonnage de l'estimateur Horvitz-Thompson* $\hat{t}_{y\pi} - t_y$ *peut se décomposer de la façon suivante*

$$\hat{t}_{y\pi} - t_y = \sum_{i \in S} B_i^{HT}(I_i = 1) + \sum_{i \in U \setminus S} B_i^{HT}(I_i = 0) \tag{2.10.5}$$

si

$$\sum_{i \in U} (I_i - \pi_i)\, a_i = 0, \tag{2.10.6}$$

où $a_i = (1 - \pi_i)^{-1} \left\{ B_i^{HT}(I_i = 1) - (d_i - 1)\, y_i \right\}$.

Remark 2.5. On peut montrer que la condition (2.10.6) est vérifiée si le coefficient a_i ne dépend pas de i,en remarquant que $\sum_{i \in U} (I_i - \pi_i) = 0$ pour un plan de taille fixe.

Remarque 2.6. On peut également montrer que la décomposition (2.10.5) est valable pour un plan de sondage poissonien et qu'elle est approximativement respectée pour un plan de sondage stratifié aléatoire sans remise ou un plan de sondage à grand entropie de taille fixe.

Dans le cas où la décomposition (2.10.5) est valable, le biais conditionnel peut être vu comme la contibution de l'unité i a l'erreur d'échantillonnage.

Le concept de configuration est une notion centrale dans la mesure où une unité est influente dans une configuration donnée ; c'est-à-dire qu'une unité est influente pour un plan, un paramètre et un estimateur donnés ou pour un modèle (m), un paramètre et un estimateur donnés. Nous allons donner dans la suite quelques exemples de configuration pour une approche sous le plan et mettre en évidence

des exemples d'unités influentes caractérisées à l'aide du biais conditionnel. Voici quelques exemples classiques de configuration auxquels on peut être confronté lors de la phase d'estimation dans une enquête :

$\mathcal{C}_1$: (Chiffre d'affaire, Chiffre d'affaire total, sondage aléatoire simple sans remise, estimateur de Horvitz-Thompson)

$\mathcal{C}_2$: (Chiffre d'affaire, Chiffre d'affaire total, Tirage poissonien, estimateur de Horvitz-Thompson)

$\mathcal{C}_3$: (Chiffre d'affaire, Chiffre d'affaire total, sondage aléatoire simple sans remise, estimateur par le ratio)

$\mathcal{C}_4$: (Chiffre d'affaire et nombre d'employés, quotient du chiffre d'affaire par le nombre d'employés, sondage aléatoire simple sans remise, estimateur par substitution)

Considérons une population de taille 5000 pour laquelle on observe les chiffres d'affaires fictifs en milliers d'euros y, rangés par ordre croissant :

$$y_1 = 0,\ y_2 = 500, y_3 = ... = y_{4999} = 500 \,\text{et}\, y_{5000} = 2000$$

Dans ce cas, la moyenne dans la population $\bar{y}_U$ est égale à 500.2.

Supposons que l'on se trouve dans une des deux configurations suivantes :

$\mathcal{C}_1$: (Chiffre d'affaire, Chiffre d'affaire total, sondage aléatoire simple sans remise, estimateur de Horvitz-Thompson)

$\mathcal{C}_2$: (Chiffre d'affaire, Chiffre d'affaire total, Tirage poissonien avec probabilités égales $\pi_k = \frac{n}{N}$, $k \in U$, estimateur de Horvitz-Thompson)

Afin de faire le lien entre le biais conditionnel et l'instabilité des estimateurs, nous rappelons dans le tableau 2.10.1, le biais conditionnel associé à une unité sélectionnée et les formules de variance pour l'estimateur Horvitz-Thompson.

	Formule de variance	Biais conditionnel de l'unité i
Sondage aléatoire simple sans remise	$Var_p(\hat{t}_{y\pi_{SASWR}}) = N^2 \frac{(1-\frac{n}{N})}{n} S^2_{y_U}$	$B_i^{HT}(I_i = 1) = \frac{N}{N-1}(\frac{N}{n} - 1)(y_i - \bar{y}_U)$
Tirage Poissonien	$Var_p(\hat{t}_{y\pi_{POISS}}) = \sum_{k \in U} \frac{(1-\pi_k)y_k^2}{\pi_k}$	$B_i^{HT}(I_i = 1) = (d_i - 1)y_i$

TABLE 2.10.1: Résumé des formules de variance et du biais conditionnel pour l'estimateur de Horvitz-Thompson

Dans le cas d'un sondage aléatoire simple sans remise, la première unité dont le chiffre d'affaire est égale à 0 contribue fortement à la variance de l'estimateur de Horvitz-Thompson si elle est sélectionnée, alors que dans le cas poissonien, la première unité ne contribue pas à la variance de l'estimateur de Horvitz-Thompson. Ainsi, l'influence d'une unité dépend fortement du plan utilisé. On peut le voir directement pour chaque unité à l'aide du biais conditionnel : dans le premier cas, le biais conditionnel est très élevé puisque la valeur 0 est très éloignée de la moyenne $\bar{y}_U = 500,2$. Alors que dans le cas du tirage poissonien à probabilités égales $\pi_k = \frac{n}{N}$, $k \in U$, le biais conditionnel est nul, car $y_1 = 0$ et donc l'influence de la première unité est très faible dans la deuxième configuration. Ce résultat, un peu contre-intuitif, donne du crédit à l'utilisation du biais conditionnel. Enfin, l'unité ayant pour valeur $y_{5000} = 2000$, est influente pour les deux plans de sondage.

2.10.3.2 Biais conditionnel pour une approche modèle

Définition 2.13. *Soient $Y_1, ..., Y_n$ un $n-$échantillon de variables aléatoires et $y_1, .., y_n$ une réalisation du $n-$échantillon, soit θ le paramètre d'intérêt et $\hat{\theta}$ un estimateur de θ, le biais conditionnel associé à l'estimateur $\hat{\theta}$ pour l'observation i est défini par :*

$$B_i^{\hat{\theta}} = \mathbb{E}\left(\hat{\theta} - \theta | \boldsymbol{I}_N,\, Y_i = y_i\right)$$

Dans le cas d'un modèle linéaire simple donnée par l'expression (2.9.2), Beaumont et al. (2013) ont montré que le biais conditionnel associé à l'estimateur BLUP donné par la formule (2.9.3) du total $t_y = \sum_{j \in U} Y_j$ est :

$$B_i^{\hat{t}^{BLUP}} = \begin{cases} (w_i - 1)(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) & , i \in s \\ -(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}), & i \in U \backslash s \end{cases}$$

où

$$w_i = 1 + \frac{\boldsymbol{x}_i}{\sigma_i^2} \left(\sum_{i \in S} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^\top}{\sigma_i^2} \right)^{-1} \left(\sum_{i \in U \backslash S} \boldsymbol{x}_i \right).$$

L'erreur de prédiction de l'estimateur BLUP se décompose comme la somme des biais conditionnels :

$$\hat{t}^{BLUP} - t_y = \sum_{i \in U} B_i^{\hat{T}^{BLUP}}.$$

Ainsi le biais conditionnel dans le cas d'une approche modèle peut s'interpréter comme la contribution de chacune des unités i de la population à l'erreur de prédiction de l'estimateur BLUP. Beaumont et Rivest (2009) ont montré que ce type de décomposition de l'erreur de prédiction était valable pour tout estimateur calé vérifiant une équation de calage de la forme $\sum_{i \in S} w_i \boldsymbol{x}_i = \sum_{i \in U} \boldsymbol{x}_i$.

Il est aussi important de noter que la variance de cette erreur de prédiction est fonction du carré des biais conditionnels :

$$Var_m \left(\hat{t}^{BLUP} - t_y \right) = E_m \left(\sum_{i \in U} B_i^{\hat{t}^{BLUP^2}} | \boldsymbol{I}_N, \, Y_i = y_i \right)$$

Ainsi, le fait de réduire le biais conditionnel d'une unité i aura pour conséquence une réduction de la variance de l'erreur de précision.

2.11 Une revue des estimateurs robustes présents dans la littérature

2.11.1 Dans un contexte de population infinie

Dans le cadre de population infinie, plusieurs articles s'intéressent à la méthode de winsorisation qui consiste à réduire à un certain seuil à déterminer, les valeurs élevées observées dans l'échantillon. Sous l'hypothèse de disposer d'un échantillon de n variables aléatoires Y_i indépendantes et identiquement distribuées selon une loi $\mathcal{F}$, d'espérance $\mu = E_{\mathcal{F}}(Y_j)$ et variance finie, Searl (1966) propose un estimateur winsorisé de la moyenne, et ainsi qu'un seuil optimal qui dépend naturellement de la queue de distribution de la loi $\mathcal{F}$ qui modélise le n-échantillon. Il développe une version non paramétrique de la moyenne empirique pour estimer la moyenne de la population μ définie par :

$$\bar{y}_R = \frac{1}{n}\sum_{i=j}^{n} \min(y_j, K), \tag{2.11.1}$$

où R correspond au seuil de la winsorisation. Pour une distribution $\mathcal{F}$ donnée et une taille d'échantillon n fixée, la valeur du seuil qui minimise l'erreur quadratique moyenne de $\bar{y}_R$ est solution de l'équation :

$$\frac{R-\mu}{n-1} = E_{\mathcal{F}}\{\max(Y-K, 0)\}. \tag{2.11.2}$$

Lorsque l'on dispose des données historiques, on peut approcher la fonction de répartition F de la loi $\mathcal{F}$ par la fonction de répartition empirique F_n, puis résoudre l'équation (2.11.2) , ce qui conduit à un bon seuil de winsorisation si la taille de l'échantillon historique est beaucoup plus grande que n. En l'absence d'information supplémentaire, on préfèrera l'estimateur winsorisé au premier ordre étudié dans l'article de Rivest (1994). Dans cet article, Rivest étudie les estimateurs de la forme (2.11.1) en choisissant le seuil R parmi les statistiques d'ordre. Il montre que choisir $R = y_{(n-1)}$ engendre une erreur quadratique plus faible que le choix

$R = y_{(n-2)}$ si la loi $\mathcal{F}$ admet un moment d'ordre 2. De plus, il montre que sous certaines conditions, l'estimateur winsorisé au premier ordre

$$\bar{y}_1 = \bar{y} - \left(y_{(n)} - y_{(n-1)}\right)/n,$$

est le meilleur estimateur winsorisé lorsque le choix du seuil se fait parmi les statistiques d'ordre. Enfin il montre que pour une distribution $\mathcal{F}$ ayant une queue de distribution à droite plus lourde que la loi exponentielle l'estimateur winsorisé au premier ordre est le plus efficace en termes d'erreur quadratique moyenne que la moyenne empirique. L'estimateur proposé par Rivest (1994) nous servira de point de comparaison pour évaluer les performances de l'estimateur robuste construit à partir du biais conditionnel dans un contexte de population infinie ; voir chapitre 5.

2.11.2 Dans un contexte de population finie

Dans un contexte de population finie avec une approche modèle, Lee (1995), puis Beaumont et Rivest (2009) donnent un panorama quasi exhaustif des estimateurs utilisables pour un sondage aléatoire simple avec remise en l'absence d'information auxiliaire. Une méthode particulièrement utilisée est la winsorisation, qui consiste à réduire à un certain seuil les valeurs trop élevées dans l'échantillon. Dans la littérature, on distingue deux types de winsorisation.

La winsorisation standard, ou encore appelée winsorisation de type I dans le cas d'un sondage aléatoire simple sans remise, consiste à réduire la valeur des unités dépassant un certain seuil en tenant compte de leur poids. Soit $\tilde{y}_i$ la valeur de la variable y pour l'unité i après winsorisation. On a

$$\tilde{y}_i = \begin{cases} y_i & \text{si } d_i y_i \leq K \\ \frac{K}{d_i} & \text{si } d_i y_i > K \end{cases} \tag{2.11.3}$$

où $K > 0$ est le seuil de winsorisation. L'estimateur winsorisé standard du total t_y est donné par

$$\hat{t}_s = \sum_{i \in S} d_i \tilde{y}_i \tag{2.11.4}$$

Une écriture alternative consiste à exprimer $\hat{t}_s$ comme une somme pondérée des valeurs initiales au moyen de poids modifiés :

$$\hat{t}_s = \sum_{i \in S} \tilde{d}_i y_i,$$

où

$$\tilde{d}_i = d_i \frac{min\left(y_i, \frac{K}{d_i}\right)}{y_i}. \tag{2.11.5}$$

Si $min\left(y_i, \frac{K}{d_i}\right) = y_i$ (c'est-à-dire que l'unité i n'est pas influente), alors $\tilde{d}_i = d_i$. Le poids d'une unité non influente n'est donc pas modifié. Par contre, le poids modifié d'une unité influente est inférieur à d_i et peut même être inférieur à 1. Il convient de noter qu'une unité affichant une valeur $y_i = 0$ ne pose pas de problème particulier puisque sa contribution au total estimé, $\hat{t}_s$, est nulle. Dans ce cas, on peut assigner une valeur arbitraire au poids modifié $\tilde{d}_i$. D'un point de vue pratique, il est peu commode d'attribuer un poids inférieur à 1 à une unité, car on souhaite qu'au minium elle se représente. C'est pourquoi, on préfère en général la winsorisation de Dalén-Tambay, appelée winsorisation de type 2 dans le cas d'un sondage aléatoire simple sans remise. On définit les valeurs de la variable d'intérêt après winsorisation par

$$\tilde{y}_i = \begin{cases} y_i & \text{si } d_i y_i \leq K \\ \frac{K}{d_i} + \frac{1}{d_i}(y_i - \frac{K}{d_i}) & \text{si } d_i y_i > K \end{cases} \tag{2.11.6}$$

Cela conduit à l'estimateur winsorisé du total t_y :

$$\hat{t}_{DT} = \sum_{i \in S} d_i \tilde{y}_i. \tag{2.11.7}$$

Comme pour $\hat{t}_s$, une écriture alternative consiste à exprimer $\hat{t}_{DT}$ comme une somme pondérée des valeurs initiales au moyen de poids modifiés :

$$\hat{t}_{DT} = \sum_{i \in S} \tilde{d}_i y_i,$$

où

$$\tilde{d}_i = 1 + (d_i - 1) \frac{min\left(y_i, \frac{K}{d_i}\right)}{y_i}. \tag{2.11.8}$$

Comme pour l'estimateur winsorisé standard, le poids d'une unité non-influente n'est pas modifié. Encore une fois, une unité affichant une valeur $y_i = 0$ ne pose pas de problème particulier puisque sa contribution au total estimé, $\hat{t}_{DT}$, est nulle. Dans ce cas, on peut assigner une valeur arbitraire au poids modifié $\tilde{d}_i$.

Le choix du seuil est très important, car il permet de réaliser le compromis biais-variance pour l'estimateur robuste. Une approche consiste à déterminer le seuil qui minimise l'erreur quadratique moyenne estimée de l'estimateur winsorisé. Par exemple, des choix optimaux de ce seuil on été proposés par Kokic et Bell (1994) dans le cas d'un sondage aléatoire simple sans remise stratifié sous l'hypothèse d'une homogénéité intra-strate pour un total et étendu dans le cas d'un estimateur par le ratio par Clark (1995). Voir également Hulliger (1995), Rivest et Hurtubise (1995) et Beaumont et Alavi (2004) pour des plans plus complexes ou des estimateurs par la régression. Malheureusement, l'estimation de l'erreur quadratique moyenne est, en général, très complexe à réaliser.

En présence d'information auxiliaire, de nombreux estimateurs robustes ont été proposés en utilisant des méthodes de M-estimation généralisées pour limiter l'impact des unités influentes. Nous détaillons le cas particulier d'un estimateur robuste du total à partir d'une méthode de M-estimation généralisée utilisant une forme Schweppe, qui consiste à seuiller les résidus normalisés de la régression. Cela conduit, voir Hampel et al. (1986), à un estimateur de la forme :

$$\hat{t}_y^{RC} = \sum_{i \in S} y_i + \sum_{i \in U \setminus S} \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}^R + \sum_{i \in S} \frac{u_i}{h_i} \psi \left\{ h_i \frac{\left(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}^R \right)}{\sqrt{v_i}} \right\}, \tag{2.11.9}$$

où $\hat{\boldsymbol{\beta}}^R$ est un estimateur robuste de $\boldsymbol{\beta}$, h_i est un poids qui peut dépendre de $\boldsymbol{x}_i$ et du poids de sondage d_i et ψ est une fonction à valeur réel bornée vérifiant $\psi(0) = 0$ et $\psi(t) = t$ lorsque t est proche de zéro.

Il est important de noter que contrairement aux estimateurs winsorisés, les

estimateurs de la forme (2.11.9) sont consistants pour un recensement, autrement dit quelque soit le choix de h_i ou de ψ, l'estimateur robuste $\hat{t}_y^{RC}$ est égal au vrai total t_y si toutes les unités sont échantillonnées. De nombreux choix pour les poids h_i ont été étudiés dans la littérature, le cas $h_i = 1$ a été traité par Chambers (1986,1997), le cas $h_i \alpha \sqrt{v_i}$ avec $v_i = Var_m\left(y_i|\boldsymbol{X}\right)$ a été étudié par Gwet et Rivest (1992) et d'autres formes de poids déduits à partir des poids de calage ont été étudiés par Beaumont et Alavi (2004). Dans la majorité des cas, la fonction ψ est la fonction de Huber. Le choix de l'estimateur $\hat{\boldsymbol{\beta}}^R$ diffère selon l'approche considérée. Dans une approche modèle, $\hat{\boldsymbol{\beta}}^R$ provient de la théorie classique des estimateurs robustes qui modélise le lien entre y et x pour la majorité des données. Il doit être résistant aux valeurs aberrantes et atteindre une efficacité élevée lorsque les erreurs du modèle suivent une loi normale, voir Hampel et al. (1986). Le deuxième terme de l'expression (2.11.9) est alors une prédiction biaisée du total de la variable y sur les unités non échantillonnées, $\sum_{i \in U \setminus S} y_i$. Le rôle du troisième terme de l'expression (2.11.9) est de compenser le biais introduit par le deuxième terme, afin d'atteindre un certain compromis biais-variance. Le détail concernant le choix des constantes d'ajustement apparaissant dans l'estimation de $\hat{\boldsymbol{\beta}}^R$ et dans la fonction ψ est donné dans Chambers (1986) et Welsh et Ronchetti (1998).

L'expression (2.11.9) peut se réécrire de la façon suivante :

$$\hat{t}_y^{RC} = \boldsymbol{t}_x^\top \hat{\boldsymbol{\beta}}^R + \sum_{i \in S} w_{r,i}\left(\hat{\boldsymbol{\beta}}^R\right)\left(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^R\right), \tag{2.11.10}$$

avec les poids

$$w_{r,i}\left(\hat{\boldsymbol{\beta}}^R\right) = w_i \frac{\psi_i^*\left\{h_i\left(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^R\right)/\sqrt{v_i}\right\}}{h_i\left(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}^R\right)/\sqrt{v_i}}$$

où w_i désigne le poids de calage non robuste de l'unité i et

$$\psi_i^*(t) = \frac{t}{w_i} + \frac{w_i - 1}{w_i}\psi(t).$$

Dans le cas d'une approche sous le plan, β est un paramètre de nuisance. On cherche alors la valeur de β qui vérifie l'équation $U(\beta) = 0$, où

$$U(\beta) = \sum_{i \in S} w_{r,i}(\beta) \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) \frac{\boldsymbol{x}_i}{v_i}.$$

On note $\hat{\boldsymbol{\beta}}^{GM}$ la solution de l'équation $U(\beta) = 0$. L'estimateur robuste associé est alors donné par :

$$\hat{t}_y^{RC} = \boldsymbol{t}_x^\top \hat{\boldsymbol{\beta}}^{GM}.$$

Les estimateurs de cette forme ont été étudiés par Gwet et Rivest (1992) ; Hulliger (1995) et Beaumont et Alavi (2004). Le choix de la tuning constante se fait alors par une minimisation de l'erreur quadratique estimée.

Toutes ces méthodes permettent de produire des estimations robustes pour une seule variable d'intérêt. Lorsque l'on s'intéresse à l'estimation simultanée de plusieurs grandeurs impliquant différentes variables d'intérêt, les relations entre les variables ne sont plus préservées. En pratique, il est assez courant d'utiliser une variable d'intérêt de référence corrélée à la majorité des autres variables d'intérêt de l'étude, puis d'utiliser les poids issus de la méthode robuste pour toutes les variables d'intérêt, voir, par exemple Brion et al. (2013) pour une application de cette méthode pour la winsorisation sur l'enquête sectorielle annuelle de l'Insee. Dans ce cas, les relations entre les variables sont préservées.

2.12 Estimation sur petits domaines

L'estimation sur les petits domaines a connu un essor important au cours de la dernière décennie. Les commanditaires d'enquête désirent des estimations de taux de chômage ou de pauvreté à des niveaux de typologies ou géographiques de plus en plus fins. On observe des tailles d'échantillons très faibles dans les domaines, car le plan de sondage initial n'est pas prévu pour ce type d'inférence sur des

zones ou domaines réduits. Dans ce cas, l'estimation direct par un estimateur de type Horvitz-Thompson devient très périlleuse en termes de variance, étant donné que celle-ci est d'ordre $\left(n_d^{-1}\right)$ où n_d est la taille de l'échantillon dans le domaine d. On a plutôt recours à des modèles. Une grande partie des rappels énoncés dans cette partie sont inspirés des livres de Rao (2003) et Chambers et Clark (2010).

Dans le cas d'une estimation sur petits domaines, on peut distinguer deux classes de modèles correspondant à deux situations différentes. La première situation correspond au cas où le modèle proposé pour la population tient aussi dans chaque petit domaine, on dit alors que l'hypothèse "synthétique" est vérifiée. La variabilité entre les petits domaines est captée par l'intermédiaire des variables auxiliaires X. Dans cette première situation, il suffit d'estimer les paramètres au niveau du modèle de superpopulation et d'utiliser ces paramètres pour prédire les totaux dans chaque domaine. La deuxième situation correspond au cas où les modèles proposés dans chacun des petits domaines sont différents, le modèle au niveau de la population est construit comme une agrégation de ces différents modèles. Dans la plupart des cas, on se contente d'ajouter un effet aléatoire propre à chaque petits domaines afin de modéliser la variabilité entre les petits domaines.

2.12.1 L'estimateur synthétique

Dans le cas d'un modèle linéaire multiple entre Y et les variables auxiliaires contenues dans la matrice résumant l'information auxiliaire $\boldsymbol{X}$, i.e :

$$\boldsymbol{y}_N = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_U,$$

où

$$E\left(\epsilon_j|\boldsymbol{x}_j\right) = 0,$$

$$Var(\epsilon_j|\boldsymbol{x}_j) = \sigma_j^2,$$

$$\epsilon_j \text{ et } \epsilon_k \text{ indépendants lorsque } j \neq k$$

où $\boldsymbol{y}_N$ désigne le vecteur contenant les N valeurs de la variable d'intérêt Y, $\boldsymbol{X}$ la matrice $N \times p$ contenant les valeurs des p variables auxiliaires pour chaque individu et $\boldsymbol{x}_i$ le i-ème ligne de la matrice X qui contient l'information auxiliaire pour l'individu i. A partir d'un estimateur $\hat{\boldsymbol{\beta}}$ du paramètre $\boldsymbol{\beta}$, on peut prédire par exemple la moyenne $\bar{y}_d$ dans le petit domaine d par :

$$\hat{\bar{y}}_d^{SYN} = \frac{1}{N_d}\left(\sum_{i \in s_d} y_i + \sum_{i \in U_d \setminus s_d} \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\right)$$

où s_d désigne les unités échantillonnées dans le petit domaine d et U_d représente l'ensemble des unités de la population appartenant au petit domaine d. Pour calculer l'estimateur synthétique, il est nécessaire de connaitre l'appartenance aux domaines de chacune des unités de la population et de disposer de l'information auxiliaire $\boldsymbol{x}_i$ pour chacune des unités de la population. Notons que l'estimateur synthétique peut être calculé pour un petit domaine ne contenant aucune unité échantillonnée, il est alors réduit à :

$$\hat{\bar{y}}_d^{SYN} = \frac{1}{N_d}\sum_{i \in U_d} \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$$

2.12.2 Les estimateurs basés sur des modèles mixtes

Dans de nombreux cas, l'information auxiliaire $\boldsymbol{x}$ disponible au niveau des domaines n'est pas suffisante. On a alors recours à des modèles plus sophistiqués, basés sur des effets aléatoires. Une première utilisation de ces modèles a vu le jour dans un contexte d'enquête agricole dans l'article de Battese et al. (1988). Leur modèle est souvent désigné par le terme anglais "Nested-error Regression model". L'idée est d'incorporer un effet aléatoire propre à chacun des domaines pour tenir compte de la spécificité de celui-ci. Des versions plus sophistiqués de

leur modèle ont été développées pour incorporer des coefficients de régressions aléatoires, décrit en statistique classique dans l'article de Dempster et al. (1981). Les estimateurs présentés précédemment reposent sur une modélisation au niveau des individus. Parfois, l'information au niveau individu n'étant pas disponible, on a alors recours a des modèles de type Fay-Herriot (Fay et al.,1979) , qui proposent des modèles équivalents au niveau du domaine. L'ensemble des résultats présentés dans la suite de la thèse sont basés sur une approche modèle au niveau individu. Ceux-ci sont généralisables sans difficulté majeure à une modélisation au niveau des domaines. Une revue des modèles mixtes utilisés pour l'estimation sur petits domaines est disponible dans Datta (2009). Des extensions avec des structures de dépendance spatio-temporelle ont également été développées, voir , par exemple, Singh et al. (2005), Pratesi et Salvati (2008), Pereira et Coelho (2010) et Marhuenda et al. (2013).

La question du calcul et de l'estimation de l'erreur quadratique moyenne des différents estimateurs sur petits domaines a été largement étudiés dans la littérature. Les cas "Nested-error Regression model" de Battese et al. (1988), de l'estimateur issu du modèle de Dempster et al. (1981) et des estimateurs issus des modèles de type Fay-Herriot (1979) sont traités dans l'article de Prasad et Rao (1990) sous des hypothèses de normalité. Des résultats similaires sont donnés dans les articles de Datta et Lahiri (2000) et Datta et al. (2005) . Davantage de détails pour l'estimation de l'erreur quadratique en tenant compte de l'estimation des composantes de la variance sont donnés dans le livre de Rao (2003).

Le cas de l'estimation dans des petits domaines à partir d'un modèle linéaire mixte généralisé (GLMM) et de l'estimation de l'erreur quadratique moyenne est détaillé par Saei et Chambers (2003). Le cas particulier des modèles logistiques mixtes et modèles logistiques mixtes multinomiales est traité respectivement dans les articles de González-Manteigea et al. (2007) et Molina et al. (2007).

Tous les estimateurs présentés dans ce paragraphe sont très sensibles à une mauvaise spécification du modèle et à la présence d'unité influente. Afin d'obtenir des estimateurs efficaces même en cas d'une mauvaise spécification du modèle ou en présence de valeurs influentes, on a recours à des méthodes d'estimations robustes.

Dans le cas d'une variable d'intérêt continue, une première forme d'estimation robuste fait appel à des modèles de régression quantile développés par Chambers et Tzavidis (2006). Des applications sur données réelles de cette méthode sont détaillées dans les articles de Tzavidis et al. (2008) et de Salvati et al. (2010). Des méthodes bayésiennes ont également été développées par Gosh et al. (2008). Sinha et Rao (2009) proposent une version robuste des estimateurs sur petits domaines en incorporant des estimations robustes des paramètres du modèle mixte. Chambers (2013) remarque que l'estimateur proposé par Sinha et Rao (2009) peut souffrir d'un biais important et propose une correction du biais à partir d'une approche similaire à Welsh et Ronchetti (1998). L'estimateur proposé par Chambers est corrigé pour le biais engendré les unités appartenant au domaine d'intérêt. En d'autres termes, il ne traite pas le biais engendré par les unités qui appartiennent à un autre petit domaine, mais qui ont une influence à travers l'estimation des coefficients et des effets aléatoires du modèle mixte. Une approche permettant une correction globale du biais a été proposé par Dongmo Jiongo et al. (2013). Elle s'appuie sur l'utilisation du biais conditionnel pour détecter puis traiter les valeurs influentes.

Dans le cas de variables d'intérêt binaires ou discrètes, on a recours à des modèles logistiques mixtes ou des modèles de poissons mixtes afin de proposer des estimations dans les petits domaines. On a recours à ces méthodes lorsque l'on souhaite, par exemple, estimer des taux de chômage pour de petites zones géogra-

phiques. Dans ce contexte, une approche bayésienne a été développée par Maiti (2001) et des estimateurs issus de la régression quantile développés par Chambers et Tzavidis (2006) ont été étendus pour le modèle binomial mixte et pour le modèle de Poisson mixte respectivement par Chambers et al. (2014) et Tzavidis et al. (2014) . Ces derniers seront détaillés dans le chapitre 6 et serviront de point de comparaison à nos estimateurs robustes construits à partir du biais conditionnel estimé dans un modèle de type GLMM.

Nous ne détaillerons pas dans ce manuscrit l'approche bayésienne faisant par exemple appel à des modèles hiérarchiques, car elle est largement détaillée dans la partie 13 du livre de Rao (2003), et nous ne l'utiliserons pas pour proposer des estimations sur petits domaines.

Bibliographie

Ardilly, P. (2006). *Les techniques de sondage*. Editions Technip.

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28–36.

Beaumont, J.-F. and Alavi, A. (2004). Robust generalized regression estimation. *Survey Methodology*, 30, 195–208.

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555–569.

Beaumont, J.-F. and Rivest, L.-P. (2009). *Dealing with outliers in survey data*. Handbook of Statistics 29: Sample Surveys: Design, Methods and Applications, Elsevier, 247-279.

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The annals of statistics*, 470–482.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 279–292.

Binder, D.A. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology*, 22, 17–22.

Binder, D. A. and Patak, Z. (1994). Use of estimating functions for estimation

from complex surveys. *Journal of the American Statistical Association*, 89, 1035–1043.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063–1069.

Chambers, R. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official statistics*, 12, 3–32.

Chambers, R. (1997). Weighting and calibration in sample survey estimation. *Conference on Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth*, 125–147.

Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 47–69

Chambers, R., and Clark, R. (2012). *An introduction to model-based survey sampling with applications*. Oxford University Press.

Chambers, R., Salvati, N. and Tzavidis, N. (2014) *M-quantile regression models for binary data in small area estimation.*

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255–268.

Chen, J. and Rao, J.N.K. (2007). Asymptotic normality under two-phase sampling designs. Statistica Sinica, 17, 1047–1064.

Clark, R.G. (1995). Winsorization methods in sample surveys. Masters thesis, Department of Statistics, Australian National University.

Dalén, J. (1987). Practical estimators of a population total which reduce the impact of large observations. R and D Report. Statistics Sweden.

Datta, G. S. (2009). Model-based approaches to small area estimation. Handbook of Statistics 29: Sample Surveys: Design, Methods and Applications, Elsevier, 251–288.

Datta, G.S., Gosh, M., Steorts, R. and Maple, J. (2011). Bayesian benchmarking with applications to small area estimation. *Test*, 20, 574–588.

Datta, G. S. and Lahiri, P. A. (200). Unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613–628.

Datta, G. S., Rao, J.N.K. and Smith, D. D. (2005) On measuring the variability of small area estimators under a basic area level model. *Biometrika*, 92, 183–196.

Demnati, A. and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 17–26.

Dempster, A. P., Rubin, D. B. and Tsutakawa, R. K. (1981) Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341–353.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey methodology*, 25, 193–204.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382.

Dongmo Jiongo, V., Haziza, D. and Duchesne, P. (2013). Controlling the bias of robust small area estimators. *Biometrika*, 100, 843–858.

Erdös, P. and Rényi, A. On the central limit theorem for samples from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 4, 49–61.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269–277.

Firth, D. and Bennett, K. (1998) Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 3–21.

Godambe, V. and Thompson, M. E. (1986) Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127–138.

Goga, C., Deville, J.-C. and Ruiz-Gazen, A. (2009) Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika*, 96, 691–709.

González-Manteiga, W., Lombardia, M. J., Molina, I., Morales, D. and Santamaria, L. (2007) Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational statistics & data analysis*, 51, 2720–2733.

Gwet, J.-P. and Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174–1182.

Hájek, J. Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 1960, 5, 361-74

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (2011) Robust statistics: the approach based on influence functions. *John Wiley & Sons*,114.

Horvitz, D. G., and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*,47, 663–685.

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79–87.

Isaki, C. T. and Fuller, W. (1982) A. Survey design under the regression superpopulation model. *Journal of the American Statistical*, 77, 89–96.

Kokic, P.N. and Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, 10, 419–435.

Krewski, D. and Rao, J.N.K (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010–1019.

Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24, 51–56.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29, 33-44

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649–673.

Maiti, T. (2001). Robust generalized linear mixed models for small area estimation. *Journal of statistical planning and inference*, 98, 225–238.

Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay–Herriot models. *Computational Statistics & Data Analysis*, 58, 308–325.

Molina, I., Saei, A. and José Lombardia, M. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 975–1000.

Montanari, G. and Ranalli, M. (2002). Asymptotically efficient generalised regression estimators. *Journal of Official Statistics*, 18, 577–590.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M. and Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: conditional bias. *Biometrika*, 86, 923–928.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M., Jimenez-Gamero, M.D. and Muñoz-Pichardo, J. (2002). Influence diagnostics in survey sampling: estimating the conditional bias. *Metrika*, 55, 209–214.

Munoz-Pichardo, J., Munoz-Garcia, J., Moreno-Rebollo, J. and Pino-Mejias, R. (1995). A new approach to influence analysis in linear models. *Sankhyā: The Indian Journal of Statistics, Series A*, 393–409.

Pereira, L. N. and Coelho, P. S. (2010) Small area estimation of mean price of habitation transaction using time-series and cross-sectional area-level models. *Journal of Applied Statistics*, 37, 651–666.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review* , 317–337.

Prasad, N. and Rao, J.N.K (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*, 85, 163–171.

Pratesi, M. and Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical methods and applications*, 17, 113–141.

Rao, J. N.K. (2003). *Small area estimation*. Wiley.

Rivest, L.-P. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika*, 81, 373–383.

Rivest, L.-P. and Hurtubise, D. (1995). On Searls Winsorized means for skewed populations. *Survey Methodology*, 21, 119–129.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. Springer.

Saei, A. and Chambers, R. (2003). Small area estimation under linear and generalized linear mixed models with time and area effects. *S3RI Methodology Working Papers M03/15*, Southampton Statistical Sciences Research Institute.

Salvati, N., Chandra, H., Ranalli, M. G. and Chambers, R. (2010). Small area estimation using a nonparametric model-based direct estimator. *Computational Statistics & Data Analysis*, 54, 2159–2171.

Searls, D. T. (1966). An estimator for a population mean which reduces the effect of large true observations. *Journal of the American Statistical Association*, 61, 1200–1204.

Sen, P. K. (1988). 12 Asymptotics in finite population sampling *Handbook of statistics*, 6, 291–331.

Singh, B. B., Shukla, G. K. and Kundu, D. (2005) Spatio-temporal models in small area estimation. Survey Methodology, 31, 183–195.

Sinha, S. K. and Rao, J.N.K. (2009) Robust small area estimation. *Canadian Journal of Statistics*, 37, 381-399 .

Skinner, C. J., Holt, D. and Smith, T. F. *(1989). Analysis of complex surveys.* Wiley.

Solon, G., Haider, S. J. and Wooldridge, J. (2014). What Are We Weighting For? *Journal of Human Resources.*

Tambay, J.-L. (1988). An integrated approach for the treatment of outliers in sub-annual surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, Virginia, 229–234.

Thompson, M.E (1997). *Theory of sample surveys.* CRC Press.

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies: cours et exercices avec solutions : [2e cycle, écoles d'ingénieurs].* Dunod.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010) Robust estimation of small-area means and quantiles. *Australian & New Zealand Journal of Statistics*, 52, 167–186 .

Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E. and Chambers, R. (2014). Robust small area prediction for counts. *Statistical methods in medical research.*

Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2008) M-quantile models with application to poverty mapping. *Statistical Methods and Applications*, 17, 393–411.

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000*). Finite population sampling and inference: a prediction approach.* Wiley.

Welsh, A. H. and Ronchetti, E. (1998) Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 413–428.

Wolter, K. (1985). *Introduction to variance estimation.* Springer.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411–414.

Chapter 3

A method of determining the winsorization threshold, with an application to domain estimation

Résumé

Dans les enquêtes auprès des entreprises, il est courant de collecter des variables économiques dont la distribution est fortement asymétrique. Dans ce contexte, la winsorisation est fréquemment utilisée afin de traiter le problème des valeurs influentes. Cette technique requiert la détermination d'une constante qui correspond au seuil à partir duquel les grandes valeurs sont réduites. Dans cet article, nous considérons une méthode de détermination de la constante qui consiste à minimiser le plus grand biais conditionnel estimé de l'échantillon. Dans le contexte de l'estimation pour des domaines, nous proposons également une méthode permettant d'assurer la cohérence entre les estimations winsorisées calculées au niveau des domaines et l'estimation winsorisée calculée au niveau de la population. Les résultats de deux études par simulation suggèrent que les méthodes proposées conduisent à des estimateurs winsorisés ayant de bonnes propriétés en termes de biais et d'efficacité relative.

Mots clés : Biais conditionnel ; Estimation robuste ; Estimateur winsorisé ; Valeurs influentes.

Abstract

In business surveys, it is not unusual to collect economic variables for which the distribution is highly skewed. In this context, winsorization is often used to treat the problem of influential values. This technique requires the determination of a constant that corresponds to the threshold above which large values are reduced. In this paper, we consider a method of determining the constant which involves minimizing the sample's largest estimated conditional bias. In the context of domain estimation, we also propose a method of ensuring consistency between the domain-level winsorized estimates and the population-level winsorized estimate. The results of two simulation studies suggest that the proposed methods lead to winsorized estimators that have good bias and relative efficiency properties.

Key Words: Conditional bias; Robust estimation; Winsorized estimator; Influential values.

3.1 Introduction

In business surveys, it is not unusual to collect economic variables for which the distribution is highly skewed. In this context, we often face the problem of influential values in the sample selected. These values are typically very large, and their presence in the sample tends to make classical estimators very unstable. It is possible to guard against the impact of influential values at the design stage by systematically selecting the potentially influential units. For example, in business surveys, it is customary to use a stratified simple random sampling without-replacement design containing one or more take-all strata that are usually composed of large units. Unfortunately, it is seldom possible to completely eliminate the problem of influential values at the design stage. The strata in business surveys are usually formed using a geography variable, a size variable (for example, number of employees) and a classification variable (for example, the North American Industry Classification System (NAICS) code). In a survey that collects dozens of variables of interest, it is not unlikely that some of them will have little or no correlation with the stratification variables, which may result in the presence of influential values. This is the case in particular in Statistics Canada's environmental surveys, such as the Agricultural Water Survey, one of whose objectives is to measure the quantity of water used by Canadian farms for irrigation. It turns out that water consumption in a given year has little correlation with the stratification variables, since consumption depends in part on the weather conditions affecting the sampled farms. Another example is the Industrial Water Survey, one of whose objectives is to measure the quantity of water used. In the case of mining companies, the consumption of water for ore extraction is strongly correlated with the geophysical characteristics of the land, which are not taken into account by the stratification variables. Another problem that leads to influential values in the sample is the presence of stratum jumpers, which arises when the stratification information collected in the field is different from the in-

formation in the sampling frame. These differences are usually due to errors in the frame (for example, an outdated frame). A stratum jumper is a unit that is not in the stratum that it would have been assigned to if the information in the frame had been accurate. If a unit with a large value is assigned to a take-some stratum, it will have a large value for the variable of interest and possibly a large sampling weight, which will potentially make it very influential. In practice, it is not unusual to have between 5% and 10% stratum jumpers.

Classical estimators (such as the expansion estimator) exhibit (virtually) no bias, but they can be very unstable in presence of influential values. Robust estimators are constructed so as to limit the impact of influential values, which leads to estimators that are more stable but potentially biased. The objective is to develop robust estimation procedures whose mean square error is significantly smaller than that of classical estimators when there are influential values in the population but which do not suffer a serious loss of efficiency when there are none. The treatment of influential values usually strikes a trade-off between bias and variance. Winsorization is a method often used in business surveys to treat influential values. It involves decreasing the value and/or weight of one or more influential units to reduce their impact. Two forms of winsorization are considered: standard winsorization and the winsorization described by Dalén (1987) and Tambay (1988). These methods are described in Section 3.4. Whichever type is used, winsorization requires the determination of a constant that corresponds to the threshold above which large values are reduced. The choice of this constant is crucial, as a poor choice may lead to winsorized estimators that have a larger mean square error than classical estimators. The problem of choosing the constant has been studied by Kokic and Bell (1994) and Rivest and Hurtubise (1995), among others. In the case of a stratified simple random sampling without-replacement design, these researchers determined the constant that minimizes the estimated mean square error of the winsorized estimators. For repeated surveys, they suggest using historical data collected in previous iterations. Kokic and Bell (1994)

determined the optimal value of the constant by setting up a common mean model in each stratum and minimizing the winsorized estimator's mean square error with respect to both to the model and the sampling design. Clark (1995) generalized the results obtained by Kokic and Bell (1994) to the case of a ratio estimator and by calculating the mean square error with respect to the model only. First, we consider a different criterion, which involves finding the constant that minimizes the sample's largest estimated conditional bias. As we explain in Section 3.2, the conditional bias associated with a unit is a measure of influence that takes into account the sampling design used. The proposed method has the advantage of being simple to apply in practice. In addition, unlike the methods proposed in the literature, it does not require historical information or a model describing the distribution of the variable of interest in each stratum. Robust estimation based on the conditional bias is presented in Section 3.3. In Section 3.5, we deal with the problem of domain estimation, which is an important problem in practice. We apply a robust method separately in each domain of interest. A population-level estimator can easily be produced by aggregating the robust estimators obtained at the domain level. However, since it is defined as the sum of estimators that are all biased, the aggregate estimator could have a large bias. This point was raised by Rivest and Hidiroglou (2004). We propose a three-step approach: First, we apply a robust method separately in each domain of interest to produce initial estimates. Independently, produce an initial robust estimate at the population level. Lastly, using a method similar to calibration (e.g., Deville and Särndal 1992), modify the initial estimates so as to ensure consistency between the robust estimates obtained at the domain level and the robust estimate obtained at the population level. The problem of consistency for domains has been studied in the context of small area estimation; for example, see You, Rao and Dick (2004) and Datta, Gosh, Steorts and Maple (2011).We conclude this section with a discussion of the concept of robustness in classical statistics and robustness in finite populations. In classical statistics, we deal with infinite populations, for which

we want to estimate the mean, say. In this context, an outlier is a value that was generated under a different model from the one under which the majority of the observations were generated. The presence of outliers in the sample can be attributed to the fact that the population from which the sample is generated is a mix of distributions or that some observations are subject to measurement errors. In classical statistics, we usually want to conduct inferences about the population of inliers. The aim is therefore to construct estimators that are robust in the sense that they are not seriously affected by the presence of outliers in the sample. In this context, it is desirable to construct robust estimators that have a high breakdown point and/or a bounded influence function. In finite populations, measurement errors are corrected at the verification stage, and it is assumed that there are none left at the estimation stage. The aim is to conduct an inference about the "total" population, which includes both outliers and inliers. In other words, in contrast to classical statistics, we are not just interested in the population of inliers. In this context, estimators that have a high breakdown point and/or a bounded influence function are generally not appropriate because they can lead to large biases. We will give preference to estimators that are robust in the sense that (i) they are more stable than classical estimators when there are influential values present and almost as efficient as classical estimators when there are no influential values present, and (ii) they converge on classical estimators as the sample size and the population size increase. Simulation studies are presented in Section 3.6. Section 3.7 concludes with a discussion.

3.2 Measure of influence: Conditional bias

Consider a finite population of individuals, denoted by U, of size N. We want to estimate the total for the variable of interest y, denoted by $t = \sum_{i \in U} y_i$. From the population we select a sample S, of expected size n, using the sampling design $p(S)$. A classical estimator of t is the expansion estimator, also known as the

Horvitz-Thompson estimator, $\hat{t} = \sum_{i \in S} d_i y_i$, where $d_i = 1/\pi_i$ is the sampling weight of unit i and π_i denotes its probability of inclusion in the sample. Although the expansion estimator, $\hat{t}$, is design-unbiased for t , it can be highly unstable in the presence of influential values.

To measure the impact (or influence) that a sampled unit has on the expansion estimator, we use the concept of conditional bias of unit; see Moreno-Rebollo, Muñoz-Reyez and Muñoz-Pichardo (1999), Moreno-Rebollo, Muñoz-Reyez, Jimenez-Gamero and Muñoz-Pichardo (2002) and Beaumont, Haziza and Ruiz-Gazen (2013). Let I_i be the sample selection indicator variable for unit i such that $I_i = 1$ if $i \in S$ and $I_i = 0$, otherwise. The conditional bias of the estimator $\hat{t}$ associated with a sampled unit is defined as

$$B_{1i}^{HT} = E_p(\hat{t}|I_i = 1) - t = \sum_{j \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j, \tag{3.2.1}$$

where π_{ij} is the joint probability of inclusion of units and in the sample. In general, the conditional bias (3.2.1) is unknown, since the values of the variable of interest are observed only for the sampled units. In practice, the conditional bias must be estimated. We consider the conditionally unbiased estimator (for example, see Beaumont et al. 2013):

$$\begin{aligned} \hat{B}_{1i}^{HT} &= \sum_{j \in S} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}} \right) y_j \\ &= (d_i - 1) y_i + \sum_{j \in S, j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}} \right) y_j. \end{aligned} \tag{3.2.2}$$

This estimator is conditionally unbiased in the sense that $E_p\left(\hat{B}_{1i}^{HT} \mid I_i = 1\right) = B_{1i}^{HT}$. We make the following remarks on the conditional bias and its estimator: (i) The conditional bias (3.2.1) and its estimator (3.2.2) depend on the inclusion probabilities π_i and the joint inclusion probabilities π_{ij}. In other words, the conditional bias is a measure that takes the sampling design into account. (ii) If $\pi_i = 1$ then $B_{1i}^{HT} = 0$ and, similarly, $\hat{B}_{1i}^{HT} = 0$. That is, when $\pi_i = 1$, unit i is selected in all possible samples, and consequently $E_p(\hat{t}|I_i = 1) - t = E_p(\hat{t}) - t = 0$, since

$\hat{t}$ is a design-unbiased estimator of t unit selected systematically in the sample therefore has no influence and does not contribute to the variance of $\hat{t}$. (iii) The estimated conditional bias (3.2.2) depends on the second-order inclusion probabilities. For some designs, these probabilities may be difficult to calculate, in which case approximations will be used. For sampling designs that belong to the class of high-entropy designs (e.g., Berger 1998), a number of approximations of the second-order inclusion probabilities have been proposed in the literature; for example, see Haziza, Mecatti and Rao (2008). An alternative solution is to calculate approximations of the using Monte Carlo methods; see Fattorini (2006) and Thompson and Wu (2008).

For a stratified simple random sampling design, the conditional bias (3.2.1) associated with sampled unit in stratum is given by

$$B_{1i}^{HT} = \frac{N_h}{N_h - 1}\left(\frac{N_h}{n_h} - 1\right)(y_i - \bar{y}_{Uh}), \tag{3.2.3}$$

where n_h denotes the size of the sample selected in stratum h, $\bar{y}_{Uh} = N_h^{-1}\sum_{i\in U_h} y_i$, and U_h denotes the population of units in stratum h of size N_h, $h = 1, \ldots, H$. The estimator of the conditional bias (3.2.2) reduces to

$$\hat{B}_{1i}^{HT} = \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right)(y_i - \bar{y}_{Sh}),$$

where $\bar{y}_{Sh} = n_h^{-1}\sum_{i\in S_h} y_i$ and S_h is the sample in stratum h .

For a Poisson design, the conditional bias of sampled unit i is given by

$$B_i^{HT}(I_i = 1) = (d_i - 1)y_i. \tag{3.2.4}$$

In contrast to the simple random sampling without-replacement design, the conditional bias (3.2.4) is known for all units in the sample, since it does not depend on finite population parameters.

3.3 Robust estimation based on the conditional bias

To guard against the undue influence of certain units, it is advisable to construct robust estimators of the total t, that is, estimators that reduce the impact of the most influential units. We consider a class of estimators of the form

$$\hat{t}_R = \hat{t} + \Delta, \tag{3.3.1}$$

where Δ is a particular random variable. As we will see in Section 3.4, the winsorized estimators considered can be written in form (3.3.1). As in Beaumont et al. (2013), we want to determine the value of Δ that minimizes the maximum estimated conditional bias of $\hat{t}_R$ in the sample. Formally, we are seeking the value of Δ that minimizes

$$\max_{i \in S} \left\{ |\hat{B}_{1i}^R| \right\}, \tag{3.3.2}$$

where $\hat{B}_{1i}^R$ denotes the estimated conditional bias of $\hat{t}_R$ associated with sampled unit i . This conditional bias is given by

$$\begin{aligned} B_{1i}^R &= E_p(\hat{t}_R | I_i = 1) - t \\ &= B_{1i}^{HT} + E_p(\Delta | I_i = 1) \end{aligned} \tag{3.3.3}$$

which is estimated by

$$\hat{B}_{1i}^R = \hat{B}_{1i}^{HT} + \Delta, \tag{3.3.4}$$

where $\hat{B}_{1i}^{HT}$ is a conditionally unbiased estimator of B_{1i}^{HT}. If we note that Δ is a conditionally unbiased estimator of $E_p(\Delta | I_i = 1)$, it follows that the estimator of the conditional bias (3.3.4) is conditionally unbiased for B_{1i}^R. In other words, we have $E_p \left\{ \hat{B}_{1i}^R \mid I_i = 1 \right\} = B_{1i}^R$.

Beaumont et al. (2013) demonstrated that under certain regularity conditions, estimator Δ that minimizes (3.3.2) is given by

$$\Delta_{opt} = -\frac{1}{2}(\hat{B}_{min} + \hat{B}_{max}),$$

where $\hat{B}_{min} = \min_{i \in S}(\hat{B}_{1i}^{HT})$ and $\hat{B}_{max} = \max_{i \in S}(\hat{B}_{1i}^{HT})$. Estimator (3.3.1) then becomes

$$\hat{t}_R = \hat{t} - \frac{1}{2}(\hat{B}_{min} + \hat{B}_{max}). \tag{3.3.5}$$

Beaumont et al. (2013) demonstrated that under certain regularity conditions, the estimator (3.3.5) is design-consistent; that is, $\hat{t}_R - t = O_p(Nn^{-1/2})$.

3.4 Application to winsorized estimators

Estimator (3.3.5) can be written in alternative forms, which can make it easier to implement in some cases. We consider the winsorized form. This form has been widely studied in the literature. As mentioned in 3.1, standard winsorization is distinguished from Dalén-Tambay winsorization.

Standard winsorization involves decreasing the value of units that are above a particular threshold, taking their weight into account. Let $\tilde{y}_i$ be the value of variable y for unit i after winsorization. We have

$$\tilde{y}_i = \begin{cases} y_i \text{ si } d_i y_i \leq K \\ \frac{K}{d_i} \text{ si } d_i y_i > K \end{cases} \tag{3.4.1}$$

where $K > 0$ is the winsorization threshold. The standard winsorized estimator of the total t given by

$$\begin{aligned} \hat{t}_s &= \sum_{i \in S} d_i \tilde{y}_i \\ &= \hat{t} + \Delta(K), \end{aligned} \tag{3.4.2}$$

where

$$\Delta(K) = -\sum_{i \in S} \max\left(0,\, d_i y_i - K\right).$$

Hence, the estimator (3.4.2) can be written in the form (3.3.1). An alternative is to express $\hat{t}_s$ as a weighted sum of the initial values using modified weights:

$$\hat{t}_s = \sum_{i \in S} \tilde{d}_i y_i,$$

where

$$\tilde{d}_i = d_i \frac{min\left(y_i,\, \frac{K}{d_i}\right)}{y_i}. \tag{3.4.3}$$

If $min\left(y_i,\, \frac{K}{d_i}\right) = y_i$ (that is, if unit is not influential), then $\tilde{d}_i = d_i$. Thus, the weight of a non-influential unit is not modified. In contrast, the modified weight of an influential unit is less than d_i and may even be less than 1. It is worth noting that a unit with a value of $y_i = 0$ presents no particular problems, since its contribution to the estimated total, $\hat{t}_s$, is zero. In this case, an arbitrary value can be assigned to the modified weight $\tilde{d}_i$.

In the case of Dalén-Tambay winsorization, the values of the variable of interest after winsorization are defined by

$$\tilde{y}_i = \begin{cases} y_i \text{ si } d_i y_i \leq K \\ \frac{K}{d_i} + \frac{1}{d_i}(y_i - \frac{K}{d_i}) \text{ si } d_i y_i > K \end{cases} \tag{3.4.4}$$

This leads to the winsorized estimator of the total t_y:

$$\begin{aligned} \hat{t}_{DT} &= \sum_{i \in S} d_i \tilde{y}_i. \\ &= \hat{t} + \Delta(K), \end{aligned} \tag{3.4.5}$$

where

$$\Delta(K) = -\sum_{i \in S} \frac{(d_i - 1)}{d_i} \max\left(0,\, d_i y_i - K\right).$$

Estimator (3.4.5) can also be written in the form (3.3.1). As in the case of $\hat{t}_s$, an alternative is to express $\hat{t}_{DT}$ as a weighted sum of the initial values using modified weights:

$$\hat{t}_{DT} = \sum_{i \in S} \tilde{d}_i y_i,$$

where

$$\tilde{d}_i = 1 + (d_i - 1)\frac{min\left(y_i,\ \frac{K}{d_i}\right)}{y_i}. \tag{3.4.6}$$

As in the case of the standard winsorized estimator, the weight of a non-influential unit is not modified. Unlike standard winsorization, Dalén-Tambay winsorization guarantees that the modified weights will not be less than 1. Once again, a unit with a value of $y_i = 0$ presents no particular problems, since its contribution to the estimated total, $\hat{t}_{DT}$, is zero. In this case, an arbitrary value can be assigned to the modified weight $\tilde{d}_i$.

Since the standard and Dalén-Tambay winsorized estimators are of the form (3.3.1), the optimal constant K_{opt} that minimizes (3.3.2) is obtained by solving

$$\Delta(K) = -\frac{1}{2}(\hat{B}_{min} + \hat{B}_{max})$$

or

$$\sum_{j \in S} a_j \max(0, d_j y_j - K) = \frac{\hat{B}_{min} + \hat{B}_{max}}{2}, \tag{3.4.7}$$

where $a_j = 1$ in the case of $\hat{t}_s$ and $a_j = (d_j - 1)/d_j$ in the case of $\hat{t}_{DT}$. It is shown in the Appendix that a solution to equation (3.4.7) exists under the following conditions:

1. $\pi_{ij} - \pi_i \pi_j \leq 0$; and

2. $\frac{1}{2}(\hat{B}_{min} + \hat{B}_{max}) \geq 0$.

Condition 1 is satisfied for most one-stage designs used in practice, such as stratified simple random sampling and Poisson sampling. Condition 2 implies that $\hat{t}_R$ must be less than or equal to $\hat{t}$ since by construction, a winsorized estimator cannot be greater than the Horvitz-Thompson estimator. It is generally expected that Condition 2 will be satisfied in most skewed populations encountered in business surveys and social surveys. It is also shown in the Appendix that the solution to equation (3.4.7) is unique if the above conditions are met and if $y_i \geq 0$ for $i \in S$.

The Appendix contains a brief description of an algorithm for finding the solution to equation (3.4.7).

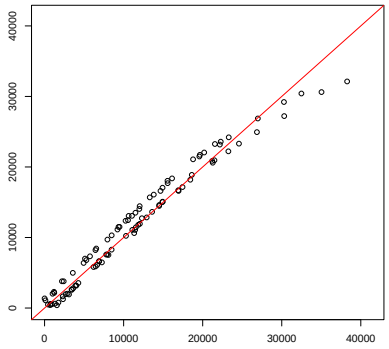
It should be noted that while the value K_{opt} is different for each type of winsorized estimator used, the resulting robust estimators are identical. In other words, we have

$$\hat{t}_s(K_{opt}) = \hat{t}_{DT}(K_{opt}) = \hat{t}_R = \hat{t} - \frac{\hat{B}_{min} + \hat{B}_{max}}{2}. \tag{3.4.8}$$

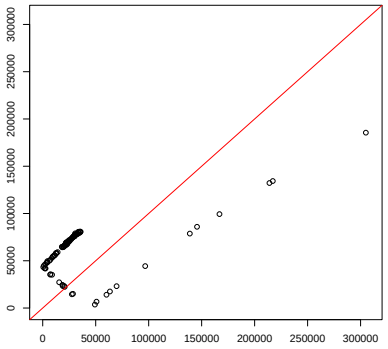
To compare the influence of each population unit with respect to the (non-robust) expansion estimator, $\hat{t}$, and its robust version (3.4.8), we carried out a simulation study. For that purpose, we generated two populations, each of size $N = 100$. One population was generated according to a normal distribution with mean 4108 and standard deviation 1500 and the other was generated according to a lognormal distribution with mean 4108 and standard deviation 7373. From each population we selected $M = 500000$samples according to two sampling designs: (i) a simple random sampling without-replacement design of size $n = 10$, and (ii) a Bernoulli design of expected size $n = 10$. First, we calculated the conditional bias of the Horvitz-Thompson estimator for a simple random sampling without-replacement design, given in (3.2.3) and for a Bernoulli design, given in (3.2.4). Note that the conditional bias of the Horvitz-Thompson estimator does not have to be approximated by simulation since all the population parameters are known. The conditional bias associated with unit i of the robust estimator given in (3.3.3) was approximated as follows: Out of the 500000 samples selected, we identified those which contained unit i. In each of these samples, we calculated the error, $\hat{t}_R - t$. Finally, we calculated the average value of $\hat{t}_R - t$ over all the samples containing unit i.

The results for the simple random sampling without-replacement design for the normal and lognormal distributions are shown in Figures 3.4.1a and 3.4.1b respectively. . The results for the Bernoulli sampling design for the normal and lognormal distributions are shown in Figures 3.4.1c and 3.4.1d respectively. In

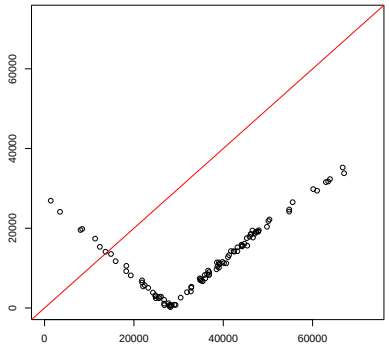
each figure, the absolute value of the conditional bias of $\hat{t}_R$ is shown in relation to the absolute value of the conditional bias of $\hat{t}$ for each population unit. The units above the first bisectrix have a conditional bias associated with $\hat{t}_R$ whose absolute value is greater than that of the conditional bias associated with $\hat{t}$. Looking first at the results for simple random sampling without replacement, we see that the behaviour of the absolute value of the conditional bias of $\hat{t}_R$ is similar to that of the absolute value of the conditional bias of $\hat{t}$, which indicates that the influence of the units is not altered significantly after robustification of the expansion estimator. This result is not surprising since the population does not contain any highly influential units. In the case of the lognormal distribution, we see that the influence of the values that have a high conditional bias associated with $\hat{t}$ has been reduced significantly. On the other hand, we note that for the majority of the data, the conditional bias of $\hat{t}_R$ is slightly higher than that of $\hat{t}$. Turning to the results for Bernoulli sampling, we see that in the case of the normal population, the influence of most units has been reduced, since the absolute value of the conditional bias of $\hat{t}_R$ is significantly lower than the absolute value of the conditional bias of $\hat{t}$. In the case of the lognormal distribution, the results are similar to those obtained with simple random sampling without replacement for the same distribution.



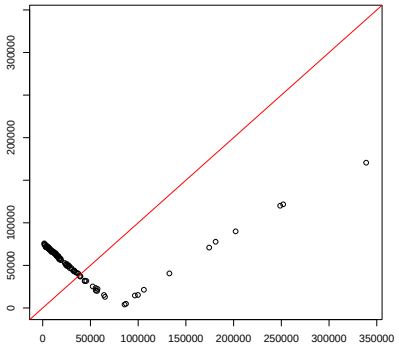
(a) Normal distribution with simple random sampling



(b) Lognormal distribution with simple random sampling



(c) Normal distribution with Bernoulli sampling



(d) Lognormal distribution with Bernoulli sampling

Figure 3.4.1: Absolute value of the conditional biases of the robust and non-robust estimators

3.5 Robust estimation of domain totals

In practice, we usually want to produce estimates for population domains as well as an estimate at the global level. Let $t_g = \sum_{i \in U_g} y_i$ be the total of y-variable in domain g. We assume that the domains form a partition of the population such that $t = \sum_{i \in U} y_i = \sum_{g=1}^{G} t_g$, where G is the number of domains. Let S_g be the set of sampled units in domain g. The expansion estimator of t_g is given by $\hat{t}_g = \sum_{i \in S_g} d_i y_i$. We have the consistency relation $\sum_{g=1}^{G} \hat{t}_g = \hat{t}$.

When there are influential values present, we can apply a robust procedure separately for each domain using the method described in 3.3, which leads to G robust estimators, $\hat{t}_{R,g}$. A robust estimator of the total at the population level, $\hat{t}_{R(agg)}$, is easily obtained by aggregating the robust estimators $\hat{t}_{R,g}$. Thus we have $\hat{t}_{R(agg)} = \sum_{g=1}^{G} \hat{t}_{R,g}$. The consistency relation between the domain-level estimates and the population-level estimate is therefore satisfied. However, aggregating G robust estimators, each suffering from a potential bias, may produce a highly biased aggregate robust estimator, $\hat{t}_{R(agg)}$. In most cases, the bias of $\hat{t}_{R(agg)}$ will be negative, since each of the $\hat{t}_{R,g}$ estimators has a negative bias.

To avoid having an estimator with an unacceptable bias, we first compute the robust estimator(3.4.8), $\hat{t}_{R,g}$, for each domain. Then we independently compute a robust estimator of the total t in the population, $\hat{t}_{R,0}$, given by (3.4.8). In this case, however, the consistency relation is no longer necessarily satisfied. In other words, we have $\hat{t}_{R,0} \neq \sum_{g=1}^{G} \hat{t}_{R,g}$, in general. It is therefore necessary to force consistency between the robust domain estimates and the aggregate robust estimate using a method similar to calibration. To do so, we compute final robust estimates $\hat{t}^{*}_{R,g}$, $g = 0, 1, .., G$, that are as close as possible to the initial robust estimates $\hat{t}_{R,g}$, based on a particular distance function, and that satisfy the calibration equation

$$\sum_{g=1}^{G} \hat{t}^{*}_{R,g} = \hat{t}^{*}_{R,0}. \tag{3.5.1}$$

In the case of the generalized chi-square distance function, we are seeking final robust estimates, $\hat{t}^{*}_{R,g}$, such that

$$\sum_{g=0}^{G} \frac{\left\{\hat{t}^{*}_{R,g} - \hat{t}_{R,g}\right\}^2}{2q_g \hat{t}_{R,g}} \tag{3.5.2}$$

is minimized subject to (3.5.1). The coefficient q_g in the above expression is a weight assigned to the initial estimate in domain g, $\hat{t}_{R,g}$, and is interpreted as its importance in the minimization problem. Using the Lagrange multipliers method, we can easily obtain a solution to this minimization problem. The solution is given

by

$$\hat{t}^*_{R,g} = \hat{t}_{R,g} - \frac{\sum_{h=0}^{G} \delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}} \delta_g q_g \hat{t}_{R,g}, \tag{3.5.3}$$

where $\delta_0 = -1$ and $\delta_g = 1$, for $g = 1, \ldots, G$.

We make the following remarks: (i) If $q_g = 0$, then the final robust estimate $\hat{t}^*_{R,g}$ is identical to the initial robust estimate $\hat{t}_{R,g}$. Thus, if we want to ensure that the initial estimate in domain g, is not modified excessively, we simply associate it with a small value of q_g. This point is also illustrated empirically in Section 3.6.2. (ii) Note that like the initial robust estimates at the domain level, $\hat{t}_{R,g}$, for $g = 1, \ldots, G$, the initial robust estimate at the population level, $\hat{t}_{R,0}$, can also be modified. (iii) If $q_0 = 0$ (in other words, the initial robust estimate for the population level is not modified) and $q_g = q$ for $g = 1, \ldots, G$, where q is a strictly positive constant, expression (3.5.3) simplifies to

$$\hat{t}^*_{R,g} = \hat{t}_{R,g} \left(\frac{\hat{t}_{R,0}}{\hat{t}_{R(agg)}} \right). \tag{3.5.4}$$

In this case, the initial estimates $\hat{t}_{R,g}$ are all modified by the same factor, $\hat{t}_{R,0}/\hat{t}_{R(agg)}$. (iv) How can we set the values of q_g in practice? It seems natural to adopt the following choice:

$$q_g = \widehat{CV}(\hat{t}_g) / \sum_{g=1}^{G} \widehat{CV}(\hat{t}_g),$$

where $\widehat{CV}(\hat{t}_g)$ is the estimated coefficient of variation (CV) associated with domain g. For example, in a repeated survey, the estimated CV observed in a previous iteration can be used. This choice of q_g is based on the fact that we will not want to make a large change in the initial estimate associated with a domain that has a small estimated CV. In such a domain, the problem of influential values is clearly less serious, and the initial robust estimate $\hat{t}_{R,g}$ is expected to be relatively close to the actual total t_g. In other words, the robust estimator $\hat{t}_{R,g}$ should have low bias and be relatively stable. It therefore makes sense not to attempt to change the initial robust estimate substantially. (v) In (3.5.2), we used the generalized chi-square distance, which leads to the linear method. In the literature on calibration

(e.g., Deville and Särndal 1992), there are a number of other calibration methods. In particular, there is the Kullback-Leibler distance, which leads to the exponential method and the logit and truncated linear methods. Using the last two methods, we can specify positive bounds C_1 and C_2 such that $C_1 \leq \hat{t}^*_{R,g}/\hat{t}_{R,g} \leq C_2$. In other words, we ensure that the ratio $\hat{t}^*_{R,g}/\hat{t}_{R,g}$, falls within the interval between C_1 and C_2. Note that the calibration procedure may lead to $\hat{t}^*_{R,g} - \hat{t}_g \geq 0$, for a certain g which is counterintuitive. In this case, we simply include the constraint $\hat{t}^*_{R,g} \leq \hat{t}_g$ for $g = 1, \ldots, G$, in the calibration procedure. (vi) An alternative is to express $\hat{t}^*_{R,g}$ as a weighted sum of the initial values using modified weights:

$$\hat{t}^*_{R,g} = \sum_{i \in S_g} \tilde{d}_i^{\,*} y_i,$$

where

$$\tilde{d}_i^{\,*} = \tilde{d}_i \left(1 - \delta_g q_g \frac{\sum_{h=0}^{G} \delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}} \right)$$

and $\tilde{d}_i$ is given by either (3.4.3) or (3.4.6). We can also write the estimator $\hat{t}^*_{R,g}$ as a weighted sum with the initial weights using modified values:

$$\hat{t}^*_{R,g} = \sum_{i \in S_g} d_i \tilde{y}_i^{\,*},$$

where

$$\tilde{y}_i^{\,*} = \tilde{y}_i \left(1 - \delta_g q_g \frac{\sum_{h=0}^{G} \delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}} \right), \; i \in g$$

and $\tilde{y}_i$ is given by either (3.4.1) or (3.4.4). (vii) We may want to find the winsorization thresholds $K_g, g = 1, \ldots, G$, such that the standard winsorized estimator or the Dalén-Tambay winsorized estimator is equal to $\hat{t}^*_{R,g}$. We can follow a procedure similar to the one in Section 3.4 and we can use an algorithm similar to the one in the Appendix. A necessary condition for the existence of a solution is that $\hat{t}_g - \hat{t}^*_{R,g} \geq 0$. (viii) With the proposed calibration procedure, more than one partition of the population can be dealt with jointly. For example, we may be interested in publishing both provincial estimates and industry estimates. If so, we

simply insert the following calibration equations into the calibration procedure:

$$\sum_{g=1}^{G} \hat{t}^*_{R,g} = \hat{t}^*_{R,0},$$
$$\sum_{l=1}^{L} \hat{t}^*_{R,l} = \hat{t}^*_{R,0},$$

where G and L denote the number of provinces and the number of industries respectively. The method can also be applied to more than two partitions of the population.

3.6 Simulation studies

3.6.1 Winsorization in a simple random sampling without-replacement design

We carried out a simulation study to examine the properties of several robust estimators using 11 populations. The first 10 of size $N = 5000$ consists of a variable of interest y. In each population, the $y-$values were generated according to the following model:

$$Y_i = U_i + \delta_i V_i,$$

where U_i, δ_i and V_i are random variables whose distributions are described in Table 3.6.1. Population 1 was generated according to a normal distribution. Populations 2 through 5 were generated using a mixture of normal distributions with contamination rates ranging from 0.5% to 5%. Populations 6 through 8 were generated according to skewed distributions. Populations 9 and 10 were generated using a mixture of lognormal distributions with contamination rates equal to 0.5% and 5%. Population 11 of size $N = 5000$ is from the information technology survey produced by the French National Institute for Statistics and Economic Studies (INSEE) in 2011. One of the survey's objectives is to estimate the e-commerce sales of French companies. We use the "sales" variable in our simulation. The distribution of y in each population is plotted in Figure 3.6.1. In addition, Table

3.6.2 presents a number of descriptive statistics for each of the populations used. For confidentiality reasons, the units for Population 11 are not shown in the plot. Similarly, there are no descriptive statistics for Population 11 in Table 3.6.2.

In each population, we selected $M = 5000$ samples according to a simple random sampling without-replacement design of size $n = 100, 300$ and 500. For each sample, we calculated expansion estimator $\hat{t}$ and the robust estimator (3.4.8). Let$y_{(1)}, \ldots, y_{(n)}$ be the values of the $y-$variable arranged in ascending order. We also calculated the once, two and three times winsorized estimators, where the $p-th$ times winsorized estimator is obtained by replacing the p largest values in the sample with the value $y_{(n-p)}$, $p = 1, 2, 3$. In a classical statistical context, Rivest (1994) showed that the once winsorized estimator has good mean-square-error properties for a large class of skewed distributions.

As a measure of the bias of an estimator $\hat{\theta}$, we calculated the Monte Carlo relative bias (in percentage):

$$BR_{MC}(\hat{\theta}) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - t\right)}{t} \times 100,$$

where $\hat{\theta}_{(m)}$ denotes the estimator $\hat{\theta}$ in sample m, $m = 1, \ldots, 5000$. We also calculated the relative efficiency of the robust estimators with respect to the expansion estimator, $\hat{t}$:

$$RE_{MC}(\hat{\theta}) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - t\right)^2}{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{t}_{(m)} - t\right)^2} \times 100.$$

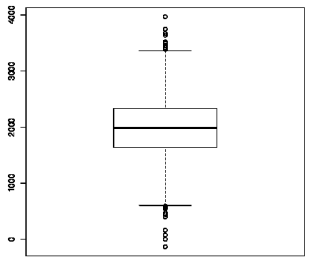
The results are shown in Table 3.6.3.

Population	U_i distribution	Mixture	δ_i distribution	V_i distribution
1	$\mathcal{N}(2000, 500)$	No		
2	$\mathcal{N}(2000, 500)$	Yes	$\mathcal{B}(0.005)$	$\mathcal{N}(50000, 10000)$
3	$\mathcal{N}(2000, 500)$	Yes	$\mathcal{B}(0.01)$	$\mathcal{N}(50000, 10000)$
4	$\mathcal{N}(2000, 500)$	Yes	$\mathcal{B}(0.02)$	$\mathcal{N}(50000, 10000)$
5	$\mathcal{N}(2000, 500)$	Yes	$\mathcal{B}(0.05)$	$\mathcal{N}(50000, 10000)$
6	$\mathcal{L}$og-$\mathcal{N}(log(2000), 1.2)$	No		
7	$\mathcal{L}$og-$\mathcal{N}(log(2000), 1.5)$	No		
8	$\mathcal{F}$rechet$(2000, 2.5, 2.1)$	No		
9	$\mathcal{L}$og-$\mathcal{N}(log(2000), 1.2)$	Yes	$\mathcal{B}(0.05)$	$\mathcal{L}$og-$\mathcal{N}(log(5000), 1.2)$
10	$\mathcal{L}$og-$\mathcal{N}(log(2000), 1.2)$	Yes	$\mathcal{B}(0.05)$	$\mathcal{L}$og-$\mathcal{N}(log(50000), 1.2)$

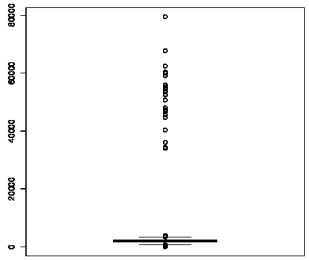
Table 3.6.1: Models used to generate the populations

Descriptive	Population									
statistic	1	2	3	4	5	6	7	8	9	10
min	132.3	314.9	105.3	275.9	187.4	23.6	7.6	2000.9	20.5	26.6
max	3968	79506	78526	80540	78690	252612	379751	2159	305612	1.3×10^6
$Q1$	1639	1667	1664	1666	1685	883	743	200	920	913
Median	1986	1993	1997	2015	2053	1996	1981	2002	2167	2041
$Q3$	2330	2337	2339	2349	2421	4505	5337	2004	5018	4927
Mean	1985	2267	2536	2976	4661	4005	6118	2004	4738	7883
Standard deviation	503	3709	5506	7119	11470	7353	17190	5.89	9796	33111
Skewness	0.0	14.0	10.2	7.3	4.3	4.2	11.6	11.8	12.1	18.4
Kurtosis	3	209	109	56	20	19	196	228	267	570
CV	0.25	1.6	2.2	2.4	2.5	1.8	2.8	2.9×10^{-3}	2.0	4.2

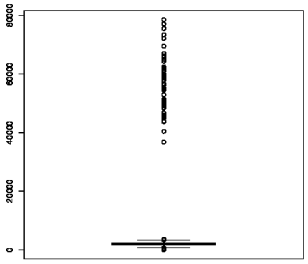
Table 3.6.2: Descriptive statistics for the 10 simulated populations



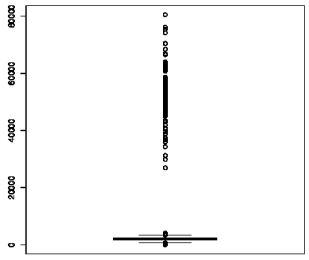
(a) Population 1



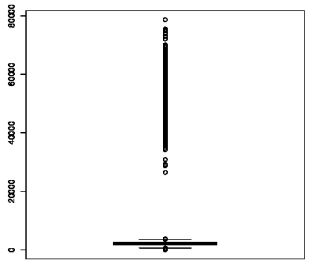
(b) Population 2



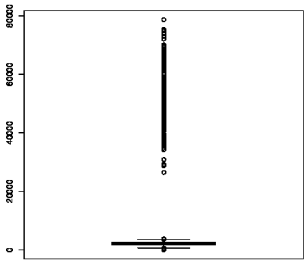
(c) Population 3



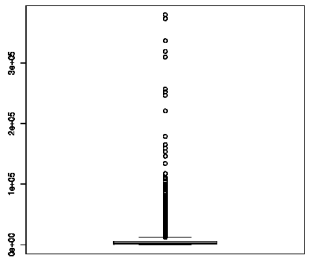
(d) Population 4



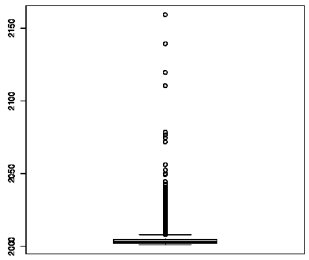
(e) Population 5



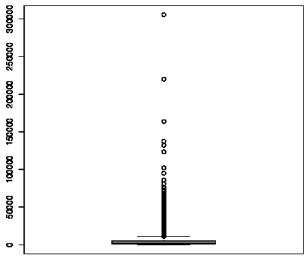
(f) Population 6



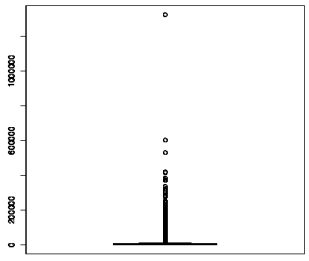
(g) Population 7



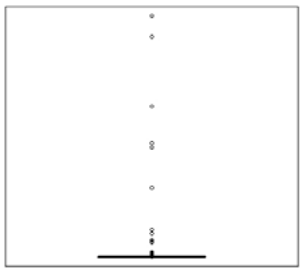
(h) Population 8



(i) Population 9



(j) Population 10



(k) Population 11

Figure 3.6.1: Distribution of the variable of interest in the 11 populations

Population	n	$\hat{t}_R$	Winsorization		
			Once	Two times	Three times
1	100	−0.1(100)	−0.1(100)	−0.2(101)	−0.3(102)
	300	0.0(100)	−0.0(100)	−0.0(100)	−0.1(100)
	500	0.0(100)	−0.0(100)	−0.0(100)	−0.0(100)
2	100	−4.9(59)	−7.5(87)	−10.7(65)	−11.9(55)
	300	−2.9(87)	−3.0(129)	−6.8(158)	−9.5(169)
	500	−1.9(96)	−1.2(122)	−3.6(175)	−6.5(226)
3	100	−6.9(74)	−8.9(122)	−16.5(119)	−20.0(107)
	300	−3.5(99)	−1.9(122)	−5.6(171)	−10.6(232)
	500	−2.4(102)	−0.9(107)	−2.2(130)	−4.5(186)
4	100	−7.6(91)	−6.2(131)	−15.5(169)	−24.4(194)
	300	−2.9(101)	−0.6(103)	−2.1(118)	−4.4(154)
	500	−2.0(102)	−0.6(102)	−1.1(101)	−1.8(108)
5	100	−5.7(102)	−1.1(104)	−4.1(126)	−9.7(173)
	300	−2.2(102)	−0.4(100)	−0.8(101)	−1.4(102)
	500	−1.2(100)	−0.1(100)	−0.3(100)	−0.5(101)
6	100	−5.7(79)	−5.4(75)	−8.2(80)	−10.6(89)
	300	−2.6(84)	−2.6(79)	−3.9(81)	−5.1(88)
	500	−2.0(86)	−2.0(81)	−3.0(82)	−3.8(88)
7	100	−8.4(72)	−9.3(73)	−14.7(72)	−18.7(79)
	300	−4.5(86)	−4.4(95)	−7.8(91)	−10.2(95)
	500	−3.5(94)	−3.1(105)	−6.0(106)	−8.1(109)
8	100	−0.0(69)	−0.0(75)	−0.0(77)	−0.0(85)
	300	−0.082)	−0.0(88)	−0.0(87)	−0.0(95)
	500	−0.0(88)	−0.0(96)	−0.0(94)	−0.0(100)
9	100	−5.7(73)	−5.8(71)	−9.5(72)	−12.4(80)
	300	−3.5(87)	−3.5(85)	−5.4(88)	−6.8(98)
	500	−2.4(88)	−2.4(88)	−3.8(90)	−4.9(97)
10	100	−13.5(68)	−15.0(70)	−24.6(76)	−31.7(89)
	300	−7.5(80)	−7.2(79)	−12.1(85)	−16.3(97)
	500	−5.3(85)	−5.1(83)	−8.4(91)	−11.4(103)
11	100	−22.8(47)	−32.6(41)	−42.0(42)	−47.7(47)
	300	−14.7(65)	−20.0(77)	−29.6(68)	−34.3(75)
	500	−11.3(76)	−14.6(96)	−24.3(90)	−29.3(97)

Table 3.6.3: Monte Carlo relative bias (in %) and relative efficiency (in parentheses) of several estimators

The results presented in Table 3.6.3 show that the once-winsorized estimator has lower bias and is generally more efficient than two times and three times winsorized estimators, which is consistent with the results obtained by Rivest (1994). It is interesting to compare the robust estimator $\hat{t}_R$ and the once-winsorized estimator. In the case of Population 1, which does not contain any influential values, we see that both estimators have low bias and are as efficient as the expansion estimator. In the case of the populations with a mixture of normal distributions (Populations 2 to 5), we observe that the once-winsorized estimator is less efficient than the robust estimator in every scenario except for Population 5 with $n = 300$. In fact, the once-winsorized estimator is less efficient than the expansion estimator in every scenario except for Population 2 with $n = 100$. The robust estimator is more efficient than the expansion estimator except in Populations 4 and 5, for which we observe values relative efficiency ranging from 91% to 102. In the case of the populations with a mixture of lognormal distributions (Populations 9 and 10), we see that the bias and efficiency performance of the once-winsorized estimator and the robust estimator is very similar in all scenarios. The same is true for the skewed populations (Populations 6 to 8), for which the two estimators produce similar results. In the case of Population 11, the robust estimator has a lower bias than the once-winsorized estimator for $n = 100$ though it is less efficient (41% versus 47%). For $n = 300$ and $n = 500$, the robust estimator has a lower bias and is significantly more efficient than the once-winsorized estimator.

3.6.2 Winsorization in a stratified simple random sampling without-replacement design

We also tested the calibration method described in Section 3.5. We generated a population of size $N = 5000$ which we divided into five strata, $U_1, \ldots, U_5$, of size $N_1, \ldots, N_5$, respectively; see Table 3.6.4 for the values of N_h. In each stratum, we generated a variable of interest y according to a lognormal distribution with parameters $\log(2,000)$ and 1.5.

From the population we selected $M = 5000$ samples according to a stratified simple random sampling without-replacement design. In stratum U_h, we selected a sample S_h of size n_h according to a simple random sampling without-replacement design; see Table 3.6.4 for the sizes n_h and the corresponding sampling fractions, $f_h = n_h/N_h$.

The objective here is to estimate the total in the population, $t = \sum_{i \in U} y_i$, and the stratum totals $t_h = \sum_{i \in U_h} y_i$, $h = 1, \ldots, H$. In other words, in our example, the strata correspond to domains of interest. Since the strata form a partition of the population, we have the consistency relation, $t = \sum_{h=1}^{H} t_h$. Similarly, the expansion estimators satisfy the consistency relation $\hat{t} = \sum_{h=1}^{H} \hat{t}_h$, where $\hat{t} = \sum_{i \in S} d_i y_i$ and $\hat{t}_h = \sum_{i \in S_h} d_i y_i$ with $d_i = N_h/n_h$ if $i \in U_h$.

For each sample, we first computed the robust estimator(3.4.8) in each stratum and aggregated the robust estimates to produce an aggregate robust estimate, $\hat{t}_{R(agg)} = \sum_{h=1}^{H} \hat{t}_{R,h}$. Independently, we computed the robust estimator (3.4.8), denoted $\hat{t}_{R,0}$, at the population level. To ensure that the consistency relation (3.5.1) was satisfied, we performed the calibration procedure described in Section 3.5 to obtain the final robust estimates $\hat{t}^*_{R,h}$, $h = 0, \ldots, 5$. We used four systems of coefficients q_h : (1) $q_0 = 0$ and $q_1 = \ldots = q_5 = 1$; (2) $q_0 = 0$ and $q_h = n_h^{-1}(1 - f_h)$, $h = 1, \ldots, 5$; (3) $q_0 = 0$ and $q_h = CV(\hat{t}_h) = \sqrt{N_h^2(1 - f_h)n_h^{-1}S_h^2}/t_h$, where $S_h^2 = (N_h - 1)^{-1} \sum_{i \in U_h} (y_i - \bar{y}_{Uh})^2$, $h = 1, \ldots, 5$; (4) $q_0 = 0$ and $q_h = \widehat{CV}(\hat{t}_h) = \sqrt{N_h^2(1 - f_h)n_h^{-1}s_h^2}/\hat{t}_h$, where $s_h^2 = (n_h - 1)^{-1} \sum_{i \in S_h} (y_i - \bar{y}_{Sh})^2$, $h = 1, \ldots, 5$; We make the following remarks on the choice of the coefficients q_h: (i) For all four systems, we assigned a weight $q_0 = 0$ to estimate $\hat{t}_{R,0}$, which is equivalent to making no change in the robust estimate at the population level. In other words, we have $\hat{t}^*_{R,0} = \hat{t}_{R,0}$. (ii) The first weighting system assigns an equal weight to all strata regardless of the sample size or sampling fraction. (iii) In the case of the second system, the coefficient q_h is a function of the sample size n_h and the sampling fraction f_h but it is independent of the intra-stratum variability S_h^2. (iv) In the third and fourth systems, the choice of q_h depends on the actual CV and

the estimated CV respectively, for the reasons mentioned in Section 3.5.

Stratum	1	2	3	4	5
N_h	2000	1500	1000	400	100
n_h	20	75	100	80	80
f_h	0.01	0.05	0.1	0.2	0.8

Table 3.6.4: Characteristics of the strata

For each robust estimator, we computed the Monte Carlo relative bias (as a percentage) and the relative efficiency (with respect to the expansion estimator); see Section 3.6.1. The results are presented in Table 3.6.5.

The results show that, the initial robust estimators $\hat{t}_{R,h}$ are biased, as expected. The bias is larger in strata with a small sampling fraction. For example, in Stratum 1, for which $f_1 = 1\%$, the relative bias of $\hat{t}_{1,h}$ is -11.9%, compared with only -1.5% in Stratum 5, for which $f_5 = 80\%$. We also note that the initial robust estimators are all more efficient than the corresponding expansion estimator, with relative efficiency values ranging from 57% to 97%. The aggregate estimator $\hat{t}_{R(agg)}$ obtained by summing the initial estimators $\hat{t}_{R,h}$, $h = 1, \ldots, 5$ shows a modest bias with a value equal to -5.7% but is more efficient than the population-level expansion estimator $\hat{t}$ with a relative efficiency of 87%.

The population-level winsorized estimator,$\hat{t}_{R,0}$ shows a small bias with a value equal to-2.8% and is significantly more efficient than the expansion estimator, with a relative efficiency of 81%. The final estimators $\hat{t}^*_{R,h}$ obtained using the system of coefficients $q_h = 1$ for $h = 1, \ldots, 5$, all have lower bias than the initial estimator $\hat{t}_{R,h}$, except for Stratum 5. This is due to the fact that we force the sum of the final estimates $\hat{t}^*_{R,h}$ to calibrate on a low-bias estimator. On the other hand, the decrease in the bias is accompanied by a slight decrease in efficiency. For example, in Stratum 4, the relative efficiency is 63% for the robust estimator $\hat{t}_{R,4}$ and 66% for the final estimator $\hat{t}^*_{R,4}$. In the case of Stratum 5, the first system

of coefficients is clearly unsuitable, since it leads to a change in the estimate for this stratum, like all the other strata, when this stratum has a very high sampling fraction of 80. In fact, for this system of coefficients, estimator $\hat{t}^*_{R,5}$ is less efficient than the expansion estimator, with a relative efficiency of 104. The second choice of coefficients q_h, which takes the sampling fraction, f_h, and the sample size n_h into account, leads to some interesting results. The final robust estimator in Stratum 1, $\hat{t}^*_{R,1}$, has an appreciably lower bias than the initial estimator $\hat{t}_{R,1}$ and the final estimator based on the first system of coefficients, at the cost of a slight loss of efficiency. For Stratum 5, the estimator $\hat{t}^*_{R,5}$ has a low bias (a relative bias of -0.8%) and the same 97% efficiency as the initial estimator $\hat{t}_{R,5}$. The third and fourth q_h weighting systems lead to similar relative bias and relative efficiency results. For Stratum 1, they lead to lower relative biases than the first weighting system, at the cost of a slight loss of efficiency. For Strata 2, 3 and 4, all four systems of coefficients exhibit similar relative bias and relative efficiency. For Stratum 5, the final estimators are virtually unbiased and no less efficient that the expansion estimator.

Global estimator		$\hat{t}_{R(agg)}$	$\hat{t}_{R,0}=\hat{t}^*_{R,0}$	$\hat{t}_{R,0}=\hat{t}^*_{R,0}$	$\hat{t}_{R,0}=\hat{t}^*_{R,0}$	$\hat{t}_{R,0}=\hat{t}^*_{R,0}$
		−5.7(87)	−2.8(81)	−2.8(81)	−2.8(81)	−2.8(81)
		$\hat{t}_{R,h}$	$\hat{t}^*_{R,h}$			
			q_h			
			1	$n_h^{-1}(1-f_h)$	$CV\left(\hat{t}_h\right)$	$\widehat{CV}\left(\hat{t}_h\right)$
Stratum	1	−11.9(57)	−9.1(60)	−0.9(67)	−5.7(62)	−6.7(64)
	2	−6.3(74)	−3.4(76)	−3.3(76)	−3.3(76)	−3.1(78)
	3	−6.0(69)	−3.1(70)	−3.8(69)	−3.2(70)	−3.2(70)
	4	−6.6(63)	−3.7(66)	−4.2(65)	−3.3(66)	−3.4(70)
	5	−1.5(97)	1.5(104)	−0.8(97)	−0.2(98)	0.1(99)

Table 3.6.5: Monte Carlo relative bias (in %) and relative efficiency (in parentheses) of the robust estimators at the global level and the stratum level

3.7 Discussion

This paper outlined a proposed method for determining the threshold for winsorized estimators. This method has the advantage of being simple to apply in practice and can be used for sampling designs with unequal probabilities. We also proposed a calibration method that satisfies a consistency relation between the domain-level winsorized estimates and a population-level winsorized estimate. Although we applied the method in the case of winsorized estimators, it can be used with any type of robust estimator.

Appendix

We want to show that there exists a solution to the equation

$$-\Delta(K) = \sum_{j \in S} a_j \max(0, d_j y_j - K) = \frac{\hat{B}_{min} + \hat{B}_{max}}{2} = \hat{t} - \hat{t}_R$$

under the conditions $\pi_{ij} - \pi_i \pi_j \leq 0$ and $\frac{1}{2}(\hat{B}_{min} + \hat{B}_{max}) \geq 0$.

First, we arrange the units in order from the smallest value of $b_i = d_i y_i, i \in S$, to the largest, so that unit 1 has the smallest value of b_i and unit n the largest value. We begin by considering the case of $\frac{1}{2}(\hat{B}_{min} + \hat{B}_{max}) = 0$. We have to solve the equation $-\Delta(K) = 0$, and we can easily see that this equation is satisfied for all $K \geq b_n$.

We now turn to the case of $\frac{1}{2}(\hat{B}_{min} + \hat{B}_{max}) > 0$. We note first that the function $-\Delta(K)$ is continuous and piecewise linear for $0 \leq K \leq b_n$. The pieces are defined by the intervals $[b_{j-1}, b_j[, j = 1, ..., n$, where $b_0 = 0$. We also note that $-\Delta(0) = \sum_{j=m}^{n} a_j b_j > 0$, where m is the smallest index such that $b_m \geq 0$. By the intermediate value theorem, there is a solution to equation (3.4.7) if we can show that

$$-\Delta(b_n) = 0 < \frac{1}{2}(\hat{B}_{min} + \hat{B}_{max}) \leq -\Delta(0) = \sum_{j=m}^{n} a_j b_j. \tag{A.1}$$

The first inequality follows directly from the condition $\frac{1}{2}(\hat{B}_{min} + \hat{B}_{max}) > 0$. To prove the second inequality, we first note that $\frac{1}{2}(\hat{B}_{min} + \hat{B}_{max}) \leq \hat{B}_{max}$. If we use the estimator of the conditional bias (3.2.2) and the condition $\pi_{ij} - \pi_i \pi_j \leq 0$, we observe that $\hat{B}_{max} \leq (d_k - 1) y_k$, index k being associated with the unit that has the largest estimated conditional bias. For the Dalén-Tambay winsorized estimator, the last inequality can be rewritten as $\hat{B}_{max} \leq a_k b_k$. It follows that $a_k b_k \leq -\Delta(0) = \sum_{j=m}^{n} a_j b_j$, which completes the proof that there is a solution to equation (3.4.7). For the standard winsorized estimator, we can also easily show that $\hat{B}_{max} \leq a_k b_k$ and therefore that a solution exists. In addition, if the $y_i, i \in S$, are all positive, the function $-\Delta(K)$ is monotonically decreasing for $0 \leq K \leq b_n$ and the solution is unique.

To find the solutionK_{opt}, we find the largest index l such that $-\Delta(b_l) \geq \frac{1}{2}(\hat{B}_{min} + \hat{B}_{max})$, for $l \leq n$. The solution can then be calculated by linear interpolation between points b_l and b_{l+1}; that is,

$$K_{opt} = b_l \frac{\Delta(b_{l+1}) - \Delta(K_{opt})}{\Delta(b_{l+1}) - \Delta(b_l)} + b_{l+1} \frac{\Delta(K_{opt}) - \Delta(b_l)}{\Delta(b_{l+1}) - \Delta(b_l)},$$

where $\Delta(K_{opt}) = -\frac{1}{2}(\hat{B}_{min} + \hat{B}_{max})$.

References

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555–569.

Clark, R.G. (1995). Winsorization methods in sample surveys. Masters thesis, Department of Statistics, Australian National University.

Dalén, J. (1987). Practical estimators of a population total which reduce the impact of large observations. R and D Report. Statistics Sweden.

Datta, G.S., Gosh, M., Steorts, R. and Maple, J. (2011). Bayesian benchmarking with applications to small area estimation. *Test*, 20, 574–588.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382.

Fattorini, L. (2006). Applying the Horvitz–Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities, *Biometrika*, 93, 269–278.

Haziza, D., Mecatti, F. and Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron*, 66, 91–108.

Kokic, P.N., and Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, 10, 419–435.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M. and Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: conditional bias. *Biometrika*, 86, 923–928.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M., Jimenez-Gamero, M.D. and Muñoz-Pichardo, J. (2002). Influence diagnostics in survey sampling: estimating the conditional bias. *Metrika*, 55, 209–214.

Rivest, L.-P. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika*, 81, 373–383.

Rivest, L.-P. and Hidiroglou, M. (2004). Outlier treatment for disaggregated estimates. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, Virginia, 4248–4256.

Rivest, L.-P. and Hurtubise, D. (1995). On Searls Winsorized means for skewed populations. *Survey Methodology*, 21, 119–129.

Tambay, J.-L. (1988). An integrated approach for the treatment of outliers in sub-annual surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, Virginia, 229–234.

Thompson, M.E. and Wu, C. (2008). Simulation-based Randomized Systematic PPS Sampling Under Substitution of Units. *Survey Methodology*, 34, 3–10.

You Y., Rao J.N.K. and Dick P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631–640.

Chapter 4

Robust inference in two-phase sampling designs with application to unit nonresponse

Résumé

Dans les enquêtes auprès des entreprises, on recueille souvent des variables économiques dont la distribution est fortement asymétrique. Par conséquent, certaines unités (par exemple, celles qui appartiennent à la queue de distribution) peuvent avoir une grande influence sur les estimateurs. Une unité influente est une unité qui, étant donnés une variable d'intérêt, un paramètre d'intérêt, un plan de sondage et un estimateur, a un impact significatif sur l'erreur due à l'échantillonnage de l'estimateur considéré. Nous allons présenter une extension des travaux de Beaumont et al. (2013) détaillé dans le cas d'un plan de sondage à une phase au cas d'un sondage à deux phases. Nous étendons la notion d'influence au cas d'un sondage à deux phases et proposons une version robuste de l'estimateur par double dilatation couramment utilisé pour un tirage à deux phases. Ensuite, nous présentons une application au traitement de la non réponse, étant donné que l'échantillon des répondants est souvent considéré comme le résultat d'une seconde phase de tirage poissonienne. Une version robuste d'estimateur calé sur de l'information auxiliaire disponible aux deux niveaux d'échantillonnage est également proposée.

Mots clés : Biais conditionnel ; unité influente ; plan à deux phases ; estimation robuste ; non réponse.

Abstract

Influential units occur frequently in surveys, especially in business surveys that collect economic variables whose distributions are highly skewed. A unit is said to be influential when its inclusion or exclusion from the sample has an important impact on the magnitude of survey statistics. We extend the results of Beaumont et al. (2013) to the case of two-phase sampling designs. We extend the concept of conditional bias attached to a unit with respect to both phases and propose a robust version of the double expansion estimator, which depends on a tuning constant. Following Beaumont et al. (2013), we determine the tuning constant which minimizes the maximum estimated conditional bias. Our results can be naturally extended to the case of unit nonresponse, since the set of respondents often being viewed as a second phase sample. A robust version of calibration estimators, based on auxiliary information available at both phases, is also constructed.

Key words: Conditional bias; influential value; two-phase sampling design; robust estimation; unit nonresponse.

4.1 Introduction

In virtually all the business surveys, one must face survey the problem of influential units. We distinguish measurement errors such as gross errors or unity errors from influential units. Errors are typically detected at the editing stage and are treated either manually or by some form of imputation. In contrast, influential units are legitimate units, which are part of the finite population under study. Influential units, which can greatly affect the quality of point estimators, may occur when the distribution of the variables being collected is highly skewed. As a result, some units may exhibit extreme values. Also, a sample unit may have a large design weight, which can make it influential, especially if it is associated with a large reported value. Finally, errors in the sampling frame create a conducive ground for the occurrence of influential units in the sample. For example, missclassification errors may cause the presence of stratum jumpers in the sample.

Before defining the concept of influence in finite population sampling, we must define that of a configuration. A configuration consists of the following elements: (1) the study variable and its distribution in the population; (2) the parameter being estimated; (3) the sampling design; (4) the point estimator and (5) the inclusion or non-inclusion in the sample.

A unit is said to be influential if, given a configuration, it has a large impact on the sampling error, $\hat{\theta}-\theta$, where θ is a finite population parameter to be estimated and $\hat{\theta}$ is an estimate of θ. A unit may be highly influential with respect to a given configuration and have no influence with respect to another configuration. In fact, modifying a single element of the configuration may have a dramatic impact on the influence of a unit.

Classical estimators such as Horvitz-Thompson type estimators or calibration estimators are sensitive to the presence of influential units. The latter do not introduce bias but they tend to make the classical estimators very unstable. Hence, it seems desirable to construct robust estimators that are less sensitive to the

presence of influential units. A first approach, which is standard practice in statistical agencies, consists of reducing the survey weight associated with units that were identified as influential; e.g., Elliott and Little (2000) and Zaslavsky et al. (2001). Sometimes, the weights of influential units is reduced to one, whereas the outstanding weight is redistributed among the other units. However, the latter approach tends to induce large biases as it can often be viewed as the limiting case of more sophisticated weight reduction procedures. Other type of methods include type I and type II winsorization, where observations exceeding a prespecified cutoff value are replaced by that cutoff value; e.g., Kokic and Bell (1994) and Chambers et al. (2000). Beaumont et al. (2013) constructed a robust estimator, using the concept of conditional bias of a unit, which can be viewed a measure of influence that accounts for the sampling design, the parameter to be estimated and the corresponding point estimator; see also Moreno-Rebollo et al. (1999, 2002).

Regardless of the approach used for curbing the influence of units identified as influential, the resulting estimators involve a tuning constant c. A suitable value for c is sometimes determined by minimizing an estimator of the mean square error of the robust estimator; e.g., Hulliger (1995), Kokic and Bell (1994) and Rivest and Hurtubise (1995). However, implementing this method may require historical information and/or some modeling assumptions. Also, minimizing the estimated mean square error of the robust estimators may prove to be complex and is often achieved at the expense of simplifying assumptions. Beaumont et al. (2013) proposed an alternative method, which consists of determining the value of c that minimizes the maximum estimated conditional bias of a unit with respect to the robust estimator. Unlike the method consisting of minimizing the estimated mean square error of the robust estimator, the method of Beaumont et al. (2013) is very simple to implement and does not require historical information nor modeling assumptions.

In this paper, we extend the results of Beaumont et al. (2013) to the case of

two-phase sampling designs, which are often used in surveys when the sampling frame contains little or no auxiliary information. In this case, it may be wise to first select a large sample in order to collect data on variables that are inexpensive to obtain and that are related to the characteristics of interest. Using the variables observed in the first-phase, an efficient sampling procedure can then be used to select a (typically small) subsample from the first-phase sample in order to collect the characteristics of interest. The theory behind inference for two-phase sampling design may also be helpful in the context of unit nonresponse since the set of respondents is often viewed as a second phase sample. The results obtained in this paper for two-phase designs can thus be readily applied in order to construct a robust version of the propensity score adjusted estimator often used in the context of unit nonresponse. To the best of our knowledge, the problem of robust estimation in the presence of unit nonresponse has not been examined in the literature.

The paper is organized as follows: in Section 4.2, we describe the theoretical set-up. In Section 3, we extend the concept of conditional bias to the case of two-phase sampling designs. Some properties of the conditional bias are also presented. Then, in Section 4, we construct a robust version of the double expansion estimator of a total based on the estimated conditional bias. The choice of the tuning constant is also discussed. The application of the proposed method to the case of unit nonresponse is presented in Section 4.5. The results of a simulation study on the performance of the robust propensity score adjusted estimator are presented in Section 4.6. In Section 4.7, we extend the proposed methods to the case of calibration estimators. The case of non-invariant two-phase designs is treated in Section 4.8. We make some final remarks in Section 4.9.

4.2 Set-up

Consider a population U of size N. We are interested in estimating the population total $Y = \sum_{i \in U} y_i$ of a characteristic of interest y. We select a sample according to a two-phase sampling design: in the first phase, a sample S_1, of size n_1, is selected from U according to a given sampling design $p(S_1)$. In the second phase, a sample, S_2, of size n_2, is selected from S_1 according to the sampling design $p(S_2|S_1)$.

We adopt the following notation: let I_{1i} be a sample selection indicator attached to unit i such that $I_{1i} = 1$ if unit i is selected in S_1 and $I_{1i} = 0$, otherwise, and let $\mathbf{I}_1 = (I_1, \cdots, I_N)^\top$. Let I_{2i} be a sample selection indicator attached to unit i such that $I_{2i} = 1$ if unit i is selected in S_2 and $I_{2i} = 0$, otherwise. Let $\pi_{1i} = P(I_{1i} = 1)$ and $\pi_{1ij} = P(I_{1i} = 1, I_{1j} = 1)$ denote the first-order and second-order probabilities in S_1. Similarly, let $\pi_{2i}(\mathbf{I}_1) = P(I_{2i} = 1 \,|\mathbf{I}_1, I_{1i} = 1)$ and $\pi_{2ij}(\mathbf{I}_1) = P(I_{2i} = 1, I_{2j} = 1 \,|\mathbf{I}_1, I_{1i} = 1, I_{1j} = 1)$ denote the first-order and second-order probabilities in S_2.

We distinguish the invariant two-phase designs from the non-invariant ones. Invariant designs are those satisfying $p(S_2|S_1) = p(S_2)$. In this case, we can write $\pi_{2i}(\mathbf{I}_1) = \pi_{2i}$ and $\pi_{2ij}(\mathbf{I}_1) = \pi_{2ij}$. In the sequel, we confine to the case of invariant two-phase designs. The extension to non-invariant designs is discussed in Section 4.8.

A basic estimator of Y is the double expansion estimator

$$\hat{Y}_{DE} = \sum_{i \in S_2} d_i^* y_i, \tag{4.2.1}$$

where $d_i^* = \pi_{1i}^{-1}\pi_{2i}^{-1}$. To study the properties of (4.2.1), we express its total error as

$$\hat{Y}_{DE} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{DE} - \hat{Y}_E), \tag{4.2.2}$$

where $\hat{Y}_E = \sum_{i \in S_1} \pi_{1i}^{-1} y_i$ denotes the expansion estimator that would have been used had the design been a single phase design. The terms $\hat{Y}_E - Y$ and $\hat{Y}_{DE} - \hat{Y}_E$ on the right hand side of (4.2.2) denote the errors due to the first phase and second phase,

respectively. Let $E_1(.)$ and $V_1(.)$ denote the expectation and variance with respect to the first phase and $E_2(.\,|\mathbf{I}_1)$ and $V_2(.\,|\mathbf{I}_1)$ denote the conditional expectation and conditional variance with respect to the second phase. Noting that $E_2(\hat{Y}_{DE}\,|\mathbf{I}_1) = \hat{Y}_E$ and $E_1(\hat{Y}_E) = Y$, it follows from (4.2.2) that $E_p(\hat{Y}_{DE}) \equiv E_1E_2(\hat{Y}_{DE}\,|\mathbf{I}_1) = Y$; that is, $\hat{Y}_{DE}$ is design-unbiased for Y. The total variance of $\hat{Y}_{DE}$ is

$$V_p(\hat{Y}_{DE}) = V_1E_2(\hat{Y}_{DE}\,|\mathbf{I}_1) + E_1V_2(\hat{Y}_{DE}\,|\mathbf{I}_1) = \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi^*_{ij}}{\pi^*_i\pi^*_j} - 1\right)y_iy_j, \quad (4.2.3)$$

where $\pi^*_i = \pi_{1i}\pi_{2i}$ and $\pi^*_{ij} = \pi_{1ij}\pi_{2ij}$.

In the presence of influential units, the estimator (4.2.1) remains design-unbiased. However, its design variance may be very large. In other words, including or excluding an influential unit from the calculations may have an important impact on the magnitude of the total error, $\hat{Y}_{DE} - Y$. An influential unit may have a large impact on the first phase error, $\hat{Y}_E - Y$, and/or on the second-phase error, $\hat{Y}_{DE} - \hat{Y}_E$. In the next section, we propose a measure of influence which measures the impact of a unit at both phases.

4.3 Measuring the influence: the conditional bias

For uni-phase sampling designs, Moreno-Rebollo et al. (1999, 2002) introduced the concept of conditional bias attached to a unit as a measure of influence; see also Beaumont et al. (2013). We extend this concept to the case of a two-phase sampling design. We distinguish between three types of units: (1) the sample units, i.e., the units for which $I_{1i} = 1$ and $I_{2i} = 1$; (2) the units selected in the first-phase sample but not in the second phase, i.e., the units for which $I_{1i} = 1$ and $I_{2i} = 0$ and (3) the non-selected units, i.e., the units for which $I_{1i} = 0$ and $I_{2i} = 0$. It is worth noting that each type of unit may have an influence on the total error. However, only the influence of the sample units (i.e., the type 1 units) can be reduced at the estimation stage. In other words, nothing can be done for the type 2 and type 3 units at this stage.

The conditional bias of $\hat{Y}_{DE}$ associated with the sample unit i is defined as

$$\begin{aligned} B_i^{DE}(I_{1i}=1, I_{2i}=1) &= E_1E_2(\hat{Y}_{DE} - Y|\mathbf{I}_1, I_{1i}=1, I_{2i}=1) \\ &= E_1(\hat{Y}_E - Y|I_{1i}=1) + E_1E_2(\hat{Y}_{DE} - \hat{Y}_E|\mathbf{I}_1, I_{1i}=1, I_{2i}=1). \end{aligned} \tag{4.3.1}$$

For an arbitrary two-phase design, we obtain

$$\begin{aligned} B_i^{DE}(I_{1i}=1, I_{2i}=1) &= \sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1\right)y_j + \sum_{j\in U}\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}\left(\frac{\pi_{2ij}}{\pi_{2i}\pi_{2j}} - 1\right)y_j \\ &= \sum_{j\in U}\left(\frac{\pi_{ij}^*}{\pi_i^*\pi_j^*} - 1\right)y_j. \end{aligned} \tag{4.3.2}$$

The first term on the right hand-side of (4.3.2) is a measure of the influence of unit i on the first-phase error $\hat{Y}_E - Y$, whereas the second term is a measure of the influence on the second-phase error, $\hat{Y}_{DE} - \hat{Y}_E$. Note that the first term on the right hand-side of (4.3.2) is equal to 0 if $\pi_{1i} = 1$, which means that unit i has no influence on the first-phase error. Similarly, the second term on the right hand-side of (4.3.2) is equal to 0 if $\pi_{2i} = 1$. These are intuitively appealing properties.

Example 4.1. For simple random sampling without replacement in both phases, (4.3.2) reduces to

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) = \frac{N}{(N-1)}\left(\frac{N}{n_2} - 1\right)(y_i - \bar{Y}),$$

where $\bar{Y} = Y/N$. The previous expression suggests that a unit has a large influence if its y-value is far from the population mean $\bar{Y}$.

Example 4.2. For Poisson sampling in both phases, (4.3.2) reduces to

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) = (d_i^* - 1)y_i.$$

Hence, a unit has a large influence if its "total weight" d_i^* is large and/or if its y-value is large.

Example 4.3. For an arbitrary design in the first phase and Poisson sampling in the second phase, (4.3.2) reduces to

$$B_i^{DE}(I_{1i}=1,I_{2i}=1)=\sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}-1\right)y_j+\pi_{1i}^{-1}\left(\pi_{2i}^{-1}-1\right)y_i. \tag{4.3.3}$$

Expression (4.3.3) will be particularly useful in the context of unit nonresponse.

As mentioned above, the type 2 and type 3 units may greatly affect the quality of the estimators though nothing can be done in order to reduce their impact at the estimation stage. Their influence may also be quantified using the concept of conditional bias.

The conditional bias of $\hat{Y}_{DE}$ associated with the sample unit i in first phase but not in second phase is defined as

$$\begin{aligned} B_i^{DE}(I_{1i}=1,I_{2i}=0) &= E_1E_2(\hat{Y}_{DE}-Y|\mathbf{I}_1,I_{1i}=1,I_{2i}=0)\\ &= \sum_{j\in U}\left\{\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}\frac{\pi_{2j}-\pi_{2ij}}{\pi_{2j}(1-\pi_{2i})}-1\right\}y_j. \end{aligned}$$

The conditional bias of $\hat{Y}_{DE}$ associated with the non-sample unit i in first phase is defined as

$$\begin{aligned} B_i^{DE}(I_{1i}=0) &= E_1E_2(\hat{Y}_{DE}-Y|\mathbf{I}_1,I_{1i}=0)\\ &= -\sum_{j\in U}\frac{(\pi_{1ij}-\pi_{1j}\pi_{1i})}{\pi_{1j}(1-\pi_{1i})}y_j+\sum_{j\in U}\frac{(\pi_{1j}-\pi_{1ij})}{\pi_{1j}(1-\pi_{1i})}\left\{\frac{\pi_{2j}-\pi_{2ij}}{\pi_{2j}(1-\pi_{2i})}-1\right\}y_j. \end{aligned}$$

Remark 4.1. For some two-phase sampling designs, the total error of $\hat{Y}_{DE}$ can be expressed as

$$\begin{aligned} \hat{Y}_E-Y=\sum_{i\in S_2}B_i^{DE}(I_{1i}=1,I_{2i}=1)+\sum_{i\in S_1\setminus S_2}B_i^{DE}(I_{1i}=1,I_{2i}=0)\\ +\sum_{i\in U\setminus S_1}B_i^{DE}(I_{1i}=0). \end{aligned} \tag{4.3.4}$$

It can be shown that (4.3.4) holds if

$$\begin{aligned} \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\left[\frac{\Delta_{1ij}}{\pi_{1j}(1-\pi_{1i})}-I_{1i}\left\{\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}\frac{\pi_{2j}-\pi_{2ij}}{\pi_{2j}(1-\pi_{2i})}-\frac{\pi_{1j}-\pi_{1ij}}{\pi_{1j}(1-\pi_{1i})}\right\}\right.\\ \left.-I_{2i}\left\{\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}\frac{\Delta_{2ij}}{\pi_{2j}\pi_{2i}(1-\pi_{2i})}\right\}\right]y_j=0, \end{aligned} \tag{4.3.5}$$

where $\Delta_{1ij} = \pi_{1ij} - \pi_{1i}\pi_{1j}$ and $\Delta_{2ij} = \pi_{2ij} - \pi_{2i}\pi_{2j}$. For Poisson sampling in both phases, the condition (4.3.5) is exactly satisfied since $\Delta_{1ij} = \Delta_{2ij} = 0$ for $i \neq j$, and

$$\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}\frac{\pi_{2j} - \pi_{2ij}}{\pi_{2j}(1-\pi_{2i})} - \frac{\pi_{1j} - \pi_{1ij}}{\pi_{1j}(1-\pi_{1i})} = 0, \quad i \neq j.$$

For simple random sampling without replacement in both phases, it can be shown that (4.3.4) holds approximately provided that the population size N is large. When (4.3.4) holds, the conditional bias associated with a unit can be interpreted as its contribution to the total error of $\hat{Y}_{DE}$.

Remark 4.2. It follows from (4.2.3) and (4.3.2) that the design-variance of $\hat{Y}_{DE}$ can be expressed as

$$V_p(\hat{Y}_{DE}) = \sum_{i \in U} B_i^{DE}(I_{1i} = 1, I_{2i} = 1) y_i.$$

The above expression suggests that, when sampled, an unit may have a significant contribution to the variance of $\hat{Y}_{DE}$ if its conditional bias $B_i^{DE}(I_{1i} = 1, I_{2i} = 1)$ is large.

Remark 4.3. In general, the conditional bias (4.3.2) is unknown as it depends on population quantities. An estimator of $B_i^{DE}(I_{1i} = 1, I_{2i} = 1)$ is given by

$$\hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) = \sum_{j \in S_2} \frac{\pi_{1i}}{\pi_{1ij}}\frac{\pi_{2i}}{\pi_{2ij}}\left(\frac{\pi_{ij}^*}{\pi_i^*\pi_j^*} - 1\right) y_j. \tag{4.3.6}$$

It is easily seen that the estimator $\hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1)$ is conditionally unbiased for $B_i^{DE}(I_{1i} = 1, I_{2i} = 1)$ in the sense that

$$E_1 E_2 \left\{ \hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) | \mathbf{I}_1, I_{1i} = 1, I_{2i} = 1 \right\} = B_i^{DE}(I_{1i} = 1, I_{2i} = 1).$$

4.4 Robustifying the double expansion estimator

Following Beaumont et al. (2013), we consider a robust version of $\hat{Y}_{DE}$:

$$\hat{Y}_{DE}^R(c) = \hat{Y}_{DE} - \sum_{i \in S_2} \left\{ \hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) \right\} + \sum_{i \in S_2} \psi_c \left\{ \hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) \right\}, \tag{4.4.1}$$

where $\psi_c(.)$ is a function, whose role consists of curbing the impact of influential units and c is a tuning constant that must be determined. We use the so-called Huber function given by $\psi_c(z) = \text{sign}(z) \times \min(|z|, c)$, where c is a positive tuning constant and $\text{sign}(z) = 1$, for $z \geq 0$, while $\text{sign}(z) = -1$, otherwise.

When $\pi_{2i} = 1$ for all $i \in S_2$ (i.e., the case of a single phase sampling design), the robust estimator (4.4.1) reduces to that proposed by Beaumont et al. (2013).

The robust estimator (4.4.1) can be alternatively written as

$$\hat{Y}_{DE}^{R}(c) = \hat{Y}_{DE} + \Delta(c), \tag{4.4.2}$$

where

$$\Delta(c) = -\sum_{i \in S_2} \left[\psi_c \left\{ \hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) \right\} - \hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) \right].$$

We are interested in determining the value $\Delta(c)$ that minimizes

$$\max_{i \in S_2} \{ \hat{B}_i^{RDE}(I_{1i} = 1, I_{2i} = 1) \}, \tag{4.4.3}$$

where $\hat{B}_i^{RDE}(I_{1i} = 1, I_{2i} = 1)$ is an estimator of the conditional bias of $\hat{Y}_{DE}^{R}$ associated with unit i. Using (4.3.2), we obtain

$$\begin{aligned} B_i^{RDE}(I_{1i} = 1, I_{2i} = 1) &= E_1 E_2 (\hat{Y}_{DE}^{R} - Y | \mathbf{I}_1, I_{1i} = 1, I_{2i} = 1) \\ &= B_i^{DE}(I_{1i} = 1, I_{2i} = 1) + E_1 E_2 \left\{ \Delta(c) | \mathbf{I}_1, I_{1i} = 1, I_{2i} = 1 \right\}. \end{aligned}$$

for $B_i^{DE}(I_{1i} = 1, I_{2i} = 1)$, the conditional bias $B_i^{RDE}(I_{1i} = 1, I_{2i} = 1)$ is generally unknown. We estimate it by

$$\hat{B}_i^{RDE}(I_{1i} = 1, I_{2i} = 1) = \hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) + \Delta(c),$$

which is conditionally unbiased for $B_i^{RDE}(I_{1i} = 1, I_{2i} = 1)$; i.e.,

$$E_1 E_2 \left\{ \hat{B}_i^{RDE}(I_{1i} = 1, I_{2i} = 1) | \mathbf{I}_1, I_{1i} = 1, I_{2i} = 1 \right\} = B_i^{RDE}(I_{1i} = 1, I_{2i} = 1).$$

It is easily seen that the value $\Delta(c)$ that minimizes (4.4.3) is given by

$$\Delta(c_{opt}) = -\frac{1}{2}(\hat{B}_{min}^{DE} + \hat{B}_{max}^{DE}), \tag{4.4.4}$$

where $\hat{B}^{DE}_{min} = \min_{i \in S_2} \left\{ \hat{B}^{DE}_i (I_{1i} = 1, I_{2i} = 1) \right\}$ and $\hat{B}^{DE}_{max} = \max_{i \in S_2} \left\{ \hat{B}^{DE}_i (I_{1i} = 1, I_{2i} = 1) \right\}$.

Using (4.4.4) in (4.4.2) leads to

$$\hat{Y}^R_{DE}(c_{opt}) = \hat{Y}_{DE} - \frac{1}{2}(\hat{B}^{DE}_{min} + \hat{B}^{DE}_{max}). \tag{4.4.5}$$

Remark 4.4. Implementation of (4.4.1) can be done by modifying either the y-values or the design weights, d^*_i of sample units. First, (4.4.1) can be expressed as

$$\hat{Y}^R_{DE}(c) = \sum_{i \in S_2} d^*_i \tilde{y}_i,$$

where

$$\tilde{y}_i = y_i - \phi_i \frac{\hat{B}^{DE}_i (I_{1i} = 1, I_{2i} = 1)}{d^*_i}$$

and

$$\phi_i = 1 - \frac{\psi_c \left\{ \hat{B}^{DE}_i (I_{1i} = 1, I_{2i} = 1) \right\}}{\hat{B}^{DE}_i (I_{1i} = 1, I_{2i} = 1)}.$$

Noting that $0 \leq \psi_c(z)/z \leq 1$ it follows that $0 \leq \phi_i \leq 1$. Therefore, if the conditional bias of unit i is small, we have $\phi_i = 0$ and $\tilde{y}_i = y_i$; i.e., the value of a non-influential unit is left unchanged. In contrast, the value of influential units is modified. Alternatively, (4.4.1) can be expressed as

$$\hat{Y}^R_{DE}(c) = \sum_{i \in S_2} \tilde{d}^*_i y_i,$$

where

$$\tilde{d}^*_i = d^*_i - \phi_i \frac{\hat{B}^{DE}_i (I_{1i} = 1, I_{2i} = 1)}{y_i}.$$

Once again, if the conditional bias of unit i is small, we have $\phi_i = 0$ and $\tilde{d}^*_i = d^*_i$; i.e., the weight of a non-influential unit is left unchanged. In contrast, the weight of influential units is modified.

Remark 4.5. This estimator (4.4.4) can be obtained without actually computing the value c_{opt} so that no iterative process is required. In practice, it may be useful to compute c_{opt} for the implementation of the robust estimator. The value c_{opt} is the solution in c to the equation

$$\Delta(c) = -\frac{1}{2}(\hat{B}^{DE}_{min} + \hat{B}^{DE}_{max}).$$

Following Beaumont et al. (2013), it can be shown that, if the Huber ψ-function is used, there always exists a solution to this equation although it is not necessarily unique.

Remark 4.6. Under mild regularity conditions, the robust estimator $\hat{Y}_{DE}^{R}(c_{opt})$ is design-consistent. That is,

$$\frac{1}{N}\left(\hat{Y}_{DE}^{R}(c_{opt}) - Y\right) = O_p(n_2^{-1/2}).$$

The proof is given in the Appendix.

4.5 Application to unit nonresponse

In this section, we consider the problem of robust estimation in the context of unit nonresponse. Two-phase sampling is useful here as it provides a framework for deriving a robust version of estimators adjusted for nonresponse. In this context, S_1 denotes the sample selected from the population, whereas S_2 denotes the random set of respondents. The quantities I_{1i} and I_{2i} denote respectively the sample selection indicator and the response indicator attached to unit i. Also, π_{1i} and π_{2i} denote respectively the inclusion probability in the sample and the response probability for unit i. We assume that the units respond independently of one another; that is $\pi_{2ij} = \pi_{2i}\pi_{2j}$ for $i \neq j$. That is, the set of respondents can be viewed as a sample that would have been selected by a Poisson sampling design with (unknown) inclusion probabilities π_{2i}. If the π_{2i}'s were known, a propensity score adjusted (PSA) estimator would be given by (4.2.1) and the conditional bias associate with a responding unit would be given (4.3.3). In practice, the response probabilities π_{2i} are unknown and must be estimated. We assume that they can be parametrically modeled by

$$\pi_{2i} = m(\boldsymbol{x}_i; \boldsymbol{\alpha}), \tag{4.5.1}$$

where $m(.)$ is a known function, $\boldsymbol{x}$ is a vector of auxiliary variables available for all the sampled units (respondents and nonrespondents) and $\boldsymbol{\alpha}$ is a vector of unknown

parameters. A special case of (4.5.1) is the logistic regression model given by

$$\pi_{2i} = \left\{1 + \exp\left(-\boldsymbol{x}_i^\top \boldsymbol{\alpha}\right)\right\}^{-1}. \tag{4.5.2}$$

An estimator of π_{2i} is given by $\hat{\pi}_{2i} = m(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}})$, where $\hat{\boldsymbol{\alpha}}$ denotes a suitable estimator of $\boldsymbol{\alpha}$ (e.g., the maximum likelihood estimator). A PSA estimator of Y is given by

$$\hat{Y}_{PSA} = \sum_{i \in S_2} \frac{1}{\pi_{1i}\hat{\pi}_{2i}} y_i. \tag{4.5.3}$$

The total error of $\hat{Y}_{PSA}$ can be expressed as

$$\hat{Y}_{PSA} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{PSA} - \hat{Y}_E). \tag{4.5.4}$$

The terms $(\hat{Y}_E - Y)$ and $(\hat{Y}_{PSA} - \hat{Y}_E)$ on the right hand side of (4.5.4) denote the sampling error and the nonresponse error, respectively.

In order to develop a robust version of $\hat{Y}_{PSA}$, we must assess the influence of each sample unit with respect to $\hat{Y}_{PSA}$. The latter being a complex function of estimated totals, the conditional bias associated with a unit is virtually untractable. To overcome this difficulty, we obtain an approximate expression of the conditional bias through a first-order Taylor expansion. A first-order approximation leads to

$$\hat{Y}_{PSA} - \hat{Y}_{PSA,lin} = O_p\left(\frac{N}{n_1}\right), \tag{4.5.5}$$

where

$$\hat{Y}_{PSA,lin} = \sum_{i \in S_1} \pi_{1i}^{-1} \left\{ k_i \pi_{1i} \pi_{2i} \mathbf{h}_i^\top \hat{\boldsymbol{\gamma}} + \frac{I_{2i}}{\pi_{2i}} \left(y_i - k_i \pi_{1i} \pi_{2i} \mathbf{h}_i^\top \hat{\boldsymbol{\gamma}}\right) \right\}$$

with $\mathbf{h}_i = \partial \left\{ logit(\pi_{2i}) \right\} / \partial \boldsymbol{\alpha}$,

$$\hat{\boldsymbol{\gamma}} = \left\{ \sum_{i \in S_1} k_i \pi_{2i} (1 - \pi_{2i}) \mathbf{h}_i \mathbf{h}_i^\top \right\}^{-1} \sum_{i \in S_1} \pi_{1i}^{-1} (1 - \pi_{2i}) \mathbf{h}_i y_i$$

and k_i is a weight attached to unit i; see Kim and Kim (2007). Commonly used values of k_i are $k_i = 1$ and $k_i = \pi_{1i}^{-1}$. Using (4.5.5) in (4.5.4), we obtain

$$\hat{Y}_{PSA} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{PSA,lin} - \hat{Y}_E) + O_p\left(\frac{N}{n_1}\right). \tag{4.5.6}$$

Ignoring the higher order terms in (4.5.6), the conditional bias of the PSA estimator associated with the responding unit i is approximated by

$$B_i^{PSA}(I_{1i}=1, I_{2i}=1) \doteq E_1E_2(\hat{Y}_{PSA,lin} - Y|\mathbf{I}_1, I_{1i}=1, I_{2i}=1).$$

In the context of unit nonresponse, $E_1(.)$ denotes the expectation with respect to the sampling design and $E_2(.|\mathbf{I}_1)$ denotes the expectation with respect to the nonresponse mechanism. After some tedious but relatively straightforward algebra, we obtain

$$\begin{aligned} B_i^{PSA}(I_{1i}=1, I_{2i}=1) &\doteq \sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}-1\right)y_j + \pi_{1i}^{-1}(\pi_{2i}^{-1}-1)\left(y_i - \mathbf{c}_i^\top\boldsymbol{\gamma}\right) \\ &- \pi_{1i}^{-1}(\pi_{2i}^{-1}-1)\mathbf{c}_i^\top\mathbf{T}^{-1}\sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}-1\right)(1-\pi_{2j})\left(y_j-\mathbf{c}_j^\top\boldsymbol{\gamma}\right)\mathbf{h}_j, \end{aligned} \tag{4.5.7}$$

where $\mathbf{c}_i = k_i\pi_{1i}\pi_{2i}\mathbf{h}_i$ and

$$\boldsymbol{\gamma} = \mathbf{T}^{-1}\sum_{i\in U}(1-\pi_{2i})\mathbf{h}_i y_i$$

with

$$\mathbf{T} = \sum_{i\in U} k_i\pi_{1i}\pi_{2i}(1-\pi_{2i})\mathbf{h}_i\mathbf{h}_i^\top.$$

A robust version of $\hat{Y}_{PSA}$ is given by

$$\hat{Y}_{PSA}^R = \hat{Y}_{PSA} - \sum_{i\in S_2}\hat{B}_i^{PSA}(I_{1i}=1, I_{2i}=1) + \sum_{i\in S_2}\psi_c\left\{\hat{B}_i^{PSA}(I_{1i}=1, I_{2i}=1)\right\},$$

where $\hat{B}_i^{PSA}(I_{1i}=1, I_{2i}=1)$ is an estimator of (4.5.7) given by

$$\begin{aligned} \hat{B}_i^{PSA}(I_{1i}=1, I_{2i}=1) = &\sum_{j\in S_2}\frac{1}{\pi_{1j}\hat{\pi}_{2j}}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}-1\right)y_j + \pi_{1i}^{-1}(\hat{\pi}_{2i}^{-1}-1)\left(y_i-\hat{\mathbf{c}}_i^\top\hat{\boldsymbol{\gamma}}_{\boldsymbol{r}}\right) \\ &-\pi_{1i}^{-1}(\hat{\pi}_{2i}^{-1}-1)\hat{\mathbf{c}}_i^\top\hat{\mathbf{T}}^{-1}\sum_{j\in S_2}\frac{1}{\pi_{1j}\hat{\pi}_{2j}}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}-1\right)(1-\hat{\pi}_{2j})\left(y_j-\hat{\mathbf{c}}_j^\top\hat{\boldsymbol{\gamma}}_{\boldsymbol{r}}\right)\hat{\mathbf{h}}_j, \end{aligned}$$

where $\hat{\pi}_{2i} = m(\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}), \hat{\mathbf{c}}_i^\top = k_i\pi_{1i}\hat{\pi}_{2i}\hat{\mathbf{h}}_i, \hat{\boldsymbol{h}}_i = \partial\left\{logit(\pi_{2i})\right\}/\partial\boldsymbol{\alpha}_{|\alpha=\hat{\alpha}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{r}} = \hat{\mathbf{T}}^{-1}\sum_{i\in S_2}{\pi_{1i}}^{-1}(\hat{\pi}_{2i}^{-1}-1)\hat{\mathbf{h}}_i y_i$ and $\hat{\mathbf{T}} = \sum_{i\in S_2} k_i(1-\hat{\pi}_{2i})\hat{\mathbf{h}}_i\hat{\boldsymbol{h}}_i^\top$.

Once again, the value of c is determined so that

$$\max_{i \in S_2} \{\hat{B}_i^{RPSA}(I_{1i} = 1, I_{2i} = 1)\}$$

is minimized, where $\hat{B}_i^{RPSA}(I_{1i} = 1, I_{2i} = 1)$ is an estimator of the conditional bias of $\hat{Y}_{PSA}^R$ associated with unit i. This leads to

$$\hat{Y}_{PSA}^R = \hat{Y}_{PSA} - \frac{1}{2}\left(\hat{B}_{min}^{PSA} + \hat{B}_{max}^{PSA}\right) \tag{4.5.8}$$

with $\hat{B}_{min}^{PSA} = \min_{i \in S_2}\left\{\hat{B}_i^{PSA}(I_{1i} = 1, I_{2i} = 1)\right\}$ and $\hat{B}_{max}^{DE} = \max_{i \in S_2}\left\{\hat{B}_i^{PSA}(I_{1i} = 1, I_{2i} = 1)\right\}$.

Remark 4.7. In practice, it is customary to partition the population into weighting adjustment cells, $U_1, \ldots, U_G$, of size $N_1, \ldots, N_G$, respectively. The response probability attached to unit i in cell g is estimated by the realized response rate within the associated cell; that is,

$$\hat{\pi}_{2i} = \hat{\pi}_{2g} = \frac{\sum_{i \in S_2 \cap U_g} \pi_{1i}^{-1}}{\sum_{i \in S_1 \cap U_g} \pi_{1i}^{-1}}, \quad \text{for } i \in U_g. \tag{4.5.9}$$

Assuming that nonresponse is uniform within cells, i.e., $\pi_{2i} = \pi_{2g}$ for $i \in U_g$, the PSA estimator (4.5.3) is asymptotically unbiased for Y.

Note that the estimated response probabilities given by (4.5.9) can alternatively be obtained by fitting the logistic model (4.5.2) with $\boldsymbol{x}_i = (\delta_{1i}, \ldots, \delta_{Gi})'$, where δ_{gi} is a class indicator such that $\delta_{gi} = 1$ if unit $i \in U_g$ and $\delta_{gi} = 0$, otherwise. Therefore, the conditional bias of the PSA estimator based on G weighting cells associated with unit i can be obtained as a special case of (4.5.7). For $i \in U_g$, it is given by

$$\begin{aligned} B_i^{PSA}(I_{1i} = 1, I_{2i} = 1) &\doteq \sum_{j \in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1\right) y_j + \pi_{1i}^{-1}\left(\pi_{2g}^{-1} - 1\right)(y_i - \bar{Y}_g) \\ &- \pi_{1i}^{-1}(\pi_{2g}^{-1} - 1)\frac{1}{N_g}\left\{\frac{1}{\pi_{1i}}(y_i - \bar{y}_{U_g})\right. \\ &+ \left.\sum_{j \in U_g, j \neq i} \frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}(y_j - \bar{Y}_g)\right\} \quad , \end{aligned} \tag{4.5.10}$$

where $\bar{Y}_g = \sum_{i \in U_g} y_i / N_g$.

4.6 Empirical Study

We conducted a simulation study to assess the performance of the robust PSA estimator based on G weighting cells, in terms of relative bias and relative efficiency. We generated three populations of size $N = 5000$, each consisting of a variable of interest y and an auxiliary variable x. First, we generated random vectors $(T_{1i}, T_{2i})^\top$ from a bivariate normal distribution with mean $(5000, 5000)^\top$ and variance-covariance matrix $10^5 \times \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$. Then, the x-values and y-values were generated according to the models presented in Table 4.6.1 : $((X_1, Y_1), ..., (X_{5000}, Y_{5000}))$ $i.i.d$ with $X_i = T_{1i} + Z_{1i}Z_{2i}$ and $Y_i = T_{2i} + Z_{3i}Z_{4i}$. Figure 4.6.1 displays the relationship between y and x for each of the population. Note that Population 1 does not contain any outlier. Population 2 contains units exhibiting large vertical residuals, whereas in Population 3, some units exhibit large vertical residuals, whereas others have large horizontal residuals.

Population	Distribution of			
	Z_{1i}	Z_{2i}	Z_{3i}	Z_{4i}
1	degenerated			
2	$\mathcal{B}(0.01)$	$\mathcal{L}$og-$\mathcal{N}(log(2000), 1.5)$	degenerated	
3	$\mathcal{B}(0.01)$	$\mathcal{L}$og-$\mathcal{N}(log(2000), 1.5)$	$\mathcal{B}(0.01)$	$\mathcal{L}$og-$\mathcal{N}(log(2000), 1.5)$

Table 4.6.1: Distributions used for generating the populations

From each population, we selected $K = 5000$ samples, of size $n = 300; 500$ according to simple random sampling without replacement. In each sample, units were assigned a response probability, π_{2i}, according to the logistic function

$$\pi_{2i} = \left[1 + \exp\left\{-1 - 0.9(x_i - \bar{X})/S_x\right\}\right]^{-1},$$

where $\bar{X} = N^{-1}\sum_{i\in U} x_i$ and $S_x = \left\{(N-1)^{-1}\sum_{i\in U}(x_i - \bar{X})^2\right\}^{1/2}$. The response indicators I_{2i} were then generated from a Bernoulli distribution with parameter π_{2i}, $i = 1, \ldots, n$. The overall response rate was set to 70%.

In each sample, we computed (i) the PSA estimator given by (4.5.3) and (ii) the robust PSA estimator given by (4.5.8). The estimated response probabilities $\hat{\pi}_{2i}$ were estimated as follows: first, a logistic regression model was fitted and estimated response probabilities were obtained. These probabilities were ordered from the smallest value to the largest value. Then, the sample was partitioned into 10 weighting cells of equal size. Finally, in each class, the response probability of unit i in class g was estimated by (4.5.9).

As a measure of bias of an estimator $\hat{Y}$, we used the Monte Carlo percent relative bias given by

$$RB_{MC}\left(\hat{Y}\right) = \frac{E_{MC}(\hat{Y}) - Y}{Y} \times 100,$$

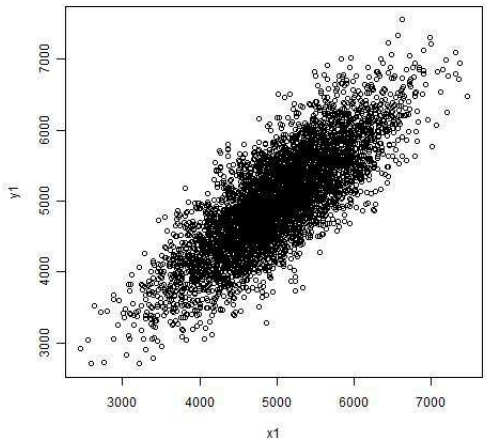
where

$$E_{MC}(\hat{Y}) = \frac{1}{K}\sum_{k=1}^{K} \hat{Y}_{(k)},$$

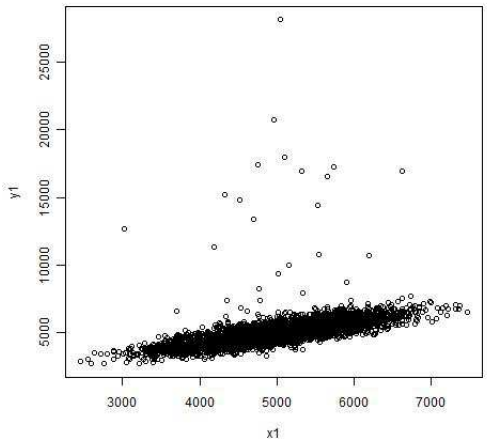
with $\hat{Y}_{(k)}$ denoting the estimator $\hat{Y}$ in the k-th sample, $k = 1, \ldots, K$. We also computed the Monte Carlo percent relative efficiency with respect to the PSA estimator

$$RE_{MC}(\hat{Y}) = 100 \times \frac{\frac{1}{K}\sum_{k=1}^{K}(\hat{Y} - Y)^2}{\frac{1}{K}\sum_{k=1}^{K}(\hat{Y}_{PSA} - Y)^2}.$$

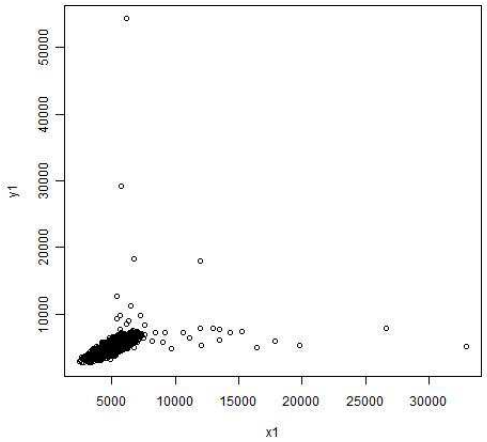
The results are shown in Table 4.6.2. From Table 4.6.2, we note that the relative bias of the PSA estimator was negligible in all the scenarios, as expected. This can be explained that the estimator was based on estimated response probabilities obtained under correct specification of the nonresponse model. The robust version of the PSA estimator also showed negligible bias in all the scenarios. In terms of relative efficiency, the robust PSA estimator was not less efficient than the PSA estimator for Population 1 with values of RE equal to 101%. For Population 2 and Population 3, the robust PSA estimator was significantly more efficient than the corresponding PSA estimator with values of RE ranging from 65% to 83%. We also note that the relative efficiency approached 100 as the expected sample size n increased, which suggest that the robust estimator is a consistent estimator of the true total Y.



(a) Population 1



(b) Population 2



(c) Population 3

Figure 4.6.1: Relationship between y and x in each population

Population	n	$\hat{\theta}_p^{RPSA}$
1	300	0.18 (101)
	500	0.17 (101)
2	300	−0.08 (78)
	500	−0.06 (83)
3	300	−0.15 (65)
	500	−0.13 (66)

Table 4.6.2: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the PSA estimator and the robust PSA estimator

4.7 Robustifying calibration estimators

In this section, the results of sections 4.2.4 are extended to the case of calibration estimators. In two-phase sampling designs, some complexity arises because auxiliary information may be available at both the population level and the first-phase level. Following Estevao and Särndal (2002, 2006), we consider two vectors of auxiliary variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of dimension J_1 and J_2. Let $\boldsymbol{x}_k = \begin{pmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{x}_{2i} \end{pmatrix}$ be the stacked vector attached to unit i of dimension $J_1 + J_2$. We assume that the auxiliary information has the following features:

- (i) the vector $\boldsymbol{x}_1$ is available for all $i \in S_2$ and the vector of population totals, $\boldsymbol{X}_1 = \sum_{i \in U} \boldsymbol{x}_{1i}$, is known.

- (ii) the vector $\boldsymbol{x}_2$ is available for all $i \in S_1$. The vector of population totals, $\boldsymbol{X}_2 = \sum_{i \in U} \boldsymbol{x}_{2i}$, is unknown but the vector of estimated totals, $\hat{\boldsymbol{X}}_2 = \sum_{i \in S_1} \pi_{1i}^{-1} \boldsymbol{x}_{2i}$, is available.

A calibration estimator of Y is given by

$$\hat{Y}_C = \sum_{i \in S_2} w_i y_i, \tag{4.7.1}$$

where the weights w_i are constructed so that the calibration constraints

$$\sum_{i\in S_2} w_i \boldsymbol{x}_i = \begin{pmatrix} \sum_{i\in U} \boldsymbol{x}_{1i} \\ \sum_{i\in S_1} \pi_i^{-1} \boldsymbol{x}_{2i} \end{pmatrix} \tag{4.7.2}$$

are satisfied. In this paper, we confine to the case of linear weighting. Extension to other types of weighting methods is relatively straightforward There exists several legitimate ways to compute the weights w_i from the auxiliary information. Here, we consider the so-called two-step calibration, top down (Estevao and Särndal, 2006). We proceed as follows: in a first step, starting with the weights, π_{1i}^{-1}, we compute intermediate weights w_{1i} for $k \in S_1$, which verifies $\sum_{k\in S_1} w_{1i}\boldsymbol{x}_{1i} = \sum_{k\in U} \boldsymbol{x}_{1i}$. In the second step, starting with the double expansion weights d_i^*, we compute the final weights w_i satisfying the calibration equations (4.7.2).

As for the PSA estimator, the calibration estimator (4.7.1) is a complex function of estimated totals. Therefore, we approximate the conditional bias of a unit through a first-order Taylor expansion, which leads to

$$\hat{Y}_C = \hat{Y}_{C,lin} + Q, \tag{4.7.3}$$

where

$$\hat{Y}_{C,lin} = \sum_{i\in S_2} d_i^* e_{2k} + \sum_{i\in S_1} \pi_{1i}^{-1} e_{1k} + \sum_{i\in U} \boldsymbol{x}_{1i}^\top \boldsymbol{\beta}^* \tag{4.7.4}$$

with

$$\boldsymbol{\beta}^* = \left(\sum_{i\in U} \boldsymbol{x}_{1i}\boldsymbol{x}_{1i}^\top\right)^{-1} \left(\sum_{i\in U} \boldsymbol{x}_{1i}\boldsymbol{x}_{1i}^\top \boldsymbol{\beta}\right), \quad \boldsymbol{\beta} = \left(\sum_{i\in U} \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1} \left(\sum_{i\in U} \boldsymbol{x}_i y_i\right)$$

and

$$e_{1i} = \boldsymbol{x}_i^\top \boldsymbol{\beta} - \boldsymbol{x}_{1i}^\top \boldsymbol{\beta}^*, \quad e_{2i} = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

The lower order term Q in (4.7.3) is given by

$$Q = -\left(\sum_{i\in S_2} d_i^* \boldsymbol{x}_k - \sum_{i\in S_1} \pi_{1i}^{-1} \boldsymbol{x}_i\right) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \left(\sum_{i\in S_1} \pi_{1i}^{-1} \boldsymbol{x}_{1i} - \sum_{i\in U} \boldsymbol{x}_{1i}\right) \left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right),$$

where

$$\hat{\boldsymbol{\beta}}^* = \left(\sum_{i\in S_1} \pi_{1i}^{-1} \boldsymbol{x}_{1i}\boldsymbol{x}_{1i}^\top\right)^{-1} \left(\sum_{i\in S_1} \pi_{1i}^{-1} \boldsymbol{x}_{1i}\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\right) \text{ and } \hat{\boldsymbol{\beta}} = \left(\sum_{i\in S_2} d_i^* \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1} \left(\sum_{i\in S_2} d_i^* \boldsymbol{x}_i y_i\right).$$

Using (4.7.4), we approximate the conditional bias of $\hat{Y}_C$ associated with unit i:

$$\begin{aligned} B_i^C(I_{1i}=1, I_{2i}=1) &= E_1E_2(\hat{Y}_C - Y|\mathbf{I}_1, I_{1i}=1, I_{2i}=1) \\ &\doteq E_1E_2(\hat{Y}_{C,lin} - Y|\mathbf{I}_1, I_{1i}=1, I_{2i}=1) \\ &\doteq \sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1\right)e_{1j} + \sum_{j\in U}\left(\frac{\pi^*_{ij}}{\pi^*_i\pi^*_j} - 1\right)e_{2j}. \end{aligned}$$

As before, the conditional bias is unknown and must be estimated. One possibility consists of estimating the latter in a nonrobust fashion. In this case, an estimator of $B_i^C(I_{1i}=1, I_{2i}=1)$ is given by

$$\hat{B}_i^C(I_{1i}=1, I_{2i}=1) = \sum_{j\in S_2}\frac{\pi_{2i}}{\pi_{2ij}}\frac{\pi_{1i}}{\pi_{1ij}}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1\right)\hat{e}_{1j} + \sum_{j\in S_2}\frac{\pi_{2i}}{\pi_{2ij}}\frac{\pi_{1i}}{\pi_{1ij}}\left(\frac{\pi^*_{ij}}{\pi^*_i\pi^*_j} - 1\right)\hat{e}_{2j},$$

where

$$\hat{e}_{1i} = \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} - \boldsymbol{x}_{1k}^\top\hat{\boldsymbol{\beta}}^* \text{ and } \hat{e}_{2k} = y_k - \boldsymbol{x}_k^\top\hat{\boldsymbol{\beta}}.$$

Following the approach of Section 4.4, we construct a robust version of $\hat{Y}_C$. Let $\hat{B}^C_{min} = \min_{i\in S_2}\left\{\hat{B}_i^C(I_{1i}=1, I_{2i}=1)\right\}$ and $\hat{B}^C_{max} = \max_{i\in S_2}\left\{\hat{B}_i^C(I_{1i}=1, I_{2i}=1)\right\}$. A robust version of $\hat{Y}_C$, denoted by $\hat{Y}_C^R$, is given by

$$\hat{Y}_C^R = \hat{Y}_C - \frac{1}{2}\left(\hat{B}^C_{max} + \hat{B}^C_{min}\right).$$

The design consistency of $\hat{Y}_C^R$ can be established using regularity conditions similar to those required for establishing the consistency of $\hat{Y}_{DE}^R$.

4.8 Non-invariant two-phase designs

In the case of non-invariant two-phase designs, the second-phase inclusion probabilities $\pi_{2i}(\mathbf{I}_1)$ and $\pi_{2ij}(\mathbf{I}_1)$ are possibly complex functions of the first-phase sample S_1. Therefore, unlike for invariant two-phase designs, the expectation with respect the first-phase sampling of the second term on the right hand-side of (4.3.2) cannot be evaluated. We have

$$E_1E_2(\hat{Y}_{DE} - \hat{Y}_E|\mathbf{I}_1, I_{1i}=1, I_{2i}=1) = E_1\left\{\sum_{j\in U}\left(\frac{\pi_{2ij}}{\pi_{2j}\pi_{2i}} - 1\right)\frac{y_j}{\pi_{1j}}I_{1j}|\mathbf{I}_1, I_{1i}=1\right\}.$$

If the y-values were available for all $i \in S_1$, a conditionally unbiased estimator of the previous expression would be

$$\sum_{j \in S_1} \left(\frac{\pi_{2ij}}{\pi_{2j}\pi_{2i}} - 1 \right) \frac{y_j}{\pi_{1j}}. \tag{4.8.1}$$

The y-values being only recorded for $i \in S_2$, the previous expression cannot be computed using the sample observations. However, we can estimate it by

$$\sum_{j \in S_2} \frac{1}{\pi_{1j}} \left(\frac{\pi_{2ij}}{\pi_{2j}\pi_{2i}} - 1 \right) \frac{\pi_{2i}}{\pi_{2ij}} y_j,$$

which is conditionally unbiased for (4.8.1). Finally, a conditionally unbiased estimator of $B_i^{DE}(I_{1i} = 1, I_{2i} = 1)$ is given by

$$\hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) = \sum_{j \in S_2} \frac{\pi_{2i}}{\pi_{2ij}} \left(\frac{\pi_{1ij}}{\pi_{1j}\pi_{1i}} - 1 \right) \frac{\pi_{1i}}{\pi_{1ij}} y_j + \sum_{j \in S_2} \frac{1}{\pi_{1j}} \left(\frac{\pi_{2ij}}{\pi_{2j}\pi_{2i}} - 1 \right) \frac{\pi_{2i}}{\pi_{2ij}} y_j. \tag{4.8.2}$$

A robust version of $\hat{Y}_{DE}$ for non-invariant two-phase deigns is given by (4.4.1), using (4.8.2) an an estimator of the conditional bias associated with unit i. If the choice of the tuning constant is obtained by minimizing the maximum estimated conditional bias of the robust estimator, then the robust version of $\hat{Y}_{DE}$ is given by (4.4.2).

4.9 Final remarks

In this paper, we have proposed a unified approach for robust estimation in two-phase sampling designs. We showed that our approach is readily applicable in the context of weighting for unit nonresponse. Estimating the mean square error of the proposed robust estimators is an important topic. The bootstrap methods seems to be attractive in this context but developing a valid bootstrap procedure for estimating the mean square error of robust estimators in the context of a two-phase sampling design is a challenging problem that requires further research.

Appendix

Design-consistency of the robust double expansion estimator

We establish the design-consistency of (4.4.5). Consistency is established by studying the decomposition

$$\hat{Y}_{DE}^{R}(c_{opt}) - Y = (\hat{Y}_{DE} - Y) - \frac{1}{2}(\hat{B}_{min}^{DE} + \hat{B}_{max}^{DE}).$$

We start by studying the term $\left(\hat{Y}_{DE} - Y\right)$ in the previous expression. For an invariant two-phase sampling design, we have

$$V_p(\hat{Y}_{DE}) \quad = \quad A_1 + A_2,$$

where

$$A_1 = \sum_{i \in U} \sum_{j \in U} \left(\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} \right) y_i y_j$$

and

$$A_2 = \sum_{i \in U} \sum_{j \in U} \left(\frac{\pi_{2ij} - \pi_{2i}\pi_{2j}}{\pi_{1i}\pi_{2i}\pi_{1j}\pi_{2j}} \right) \pi_{1ij} y_i y_j.$$

We assume that the following regularity conditions hold:

(H_1) $\lim_{N\to\infty} \frac{n_1}{N} = \alpha$ and $\lim_{N\to\infty} \frac{n_2}{n_1} = \kappa$;

(H_2) $\forall i \in U$ $\min(\pi_{1i}) > \lambda_1 > 0$ and $\min(\pi_{2i}) > \lambda_2 > 0$;

(H_3) $\forall (i,j) \in U^2$ $\min(\pi_{1ij}) > \lambda_1^* > 0$ and $\min(\pi_{2ij}) > \lambda_2^* > 0$;

(H_4) $\lim_{N\to\infty} n_1 \times \max_{i\neq j} |\pi_{1ij} - \pi_{1i}\pi_{1j}| < C_1 < \infty$ and $\lim_{N\to\infty} n_2 \times \max_{i\neq j} |\pi_{2ij} - \pi_{2i}\pi_{2j}| < C_2 < \infty$;

(H_5) $\lim_{N\to\infty}\frac{1}{N}\sum_{i\in U}y_i^2 = C_3 < \infty$. First, using $(H_1)-(H_5)$, we have

$$\begin{aligned}
|A_1| &\leq \sum_{i\in U}\sum_{j\in U}|\left(\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}}\right)y_iy_j| \\
&\leq \sum_{i\in U}|\frac{1-\pi_{1i}}{\pi_{1i}}y_i^2| + \sum_{i\in U}\sum_{j\in U, j\neq i}|\left(\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}}\right)y_iy_j| \\
&\leq \frac{1-\lambda_1}{\lambda_1}\sum_{i\in U}y_i^2 + \frac{N^2}{n_1}\frac{n_1\times\max_{i\neq j}|\pi_{1ij}-\pi_{1i}\pi_{1j}|}{\lambda_1^2}\frac{1}{N^2}\sum_{i\in U}\sum_{j\in U, j\neq i}|y_i||y_j| \\
&\leq \frac{1-\lambda_1}{\lambda_1}\sum_{i\in U}y_i^2 + \frac{N^2}{n_1}\frac{C_1}{\lambda_1^2}\left(\frac{1}{N}\sum_{i\in U}|y_i|\right)^2 \\
&\leq \frac{1-\lambda_1}{\lambda_1}N^2\frac{1}{n_1}\frac{n_1}{N}\frac{1}{N}\sum_{i\in U}y_i^2 + \frac{N^2}{n_1}\frac{C_1}{\lambda_1^2}\left(\frac{1}{N}\sum_{i\in U}|y_i|\right)^2 \\
&\leq C\frac{N^2}{n_1}.
\end{aligned}$$

It follows that $|A_1| = O(N^2n_1^{-1})$. Also,

$$\begin{aligned}
A_2 &= \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{2ij}-\pi_{2i}\pi_{2j}}{\pi_{1i}\pi_{2i}\pi_{1j}\pi_{2j}}\right)\pi_{1ij}y_iy_j \\
&= \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{2ij}-\pi_{2i}\pi_{2j}}{\pi_{1i}\pi_{2i}\pi_{1j}\pi_{2j}}\right)(\pi_{1ij}-\pi_{1i}\pi_{1j})y_iy_j + \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{2ij}-\pi_{2i}\pi_{2j}}{\pi_{2i}\pi_{2j}}\right)y_iy_j \\
&= A_3 + A_4.
\end{aligned}$$

Using similar arguments to those used for $|A_1|$, we obtain $|A_4| = O\left(\frac{N^2}{n_2}\right)$. Now,

$$\begin{aligned}
|A_3| &= |\sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{2ij}-\pi_{2i}\pi_{2j}}{\pi_{1i}\pi_{2i}\pi_{1j}\pi_{2j}}\right)(\pi_{1ij}-\pi_{1i}\pi_{1j})y_iy_j| \\
&\leq \sum_{i\in U}\frac{\pi_{2i}(1-\pi_{2i})\pi_{1i}(1-\pi_{1i})}{\pi_{1i}^2\pi_{2i}^2}y_i^2 \\
&+ \frac{N^2}{n_1n_2}\frac{n_1\times\max_{i\neq j}|\pi_{1ij}-\pi_{1i}\pi_{1j}|\times n_2\times\max_{i\neq j}|\pi_{2ij}-\pi_{2i}\pi_{2j}|}{\lambda_1^2\lambda_2^2}\times\frac{1}{N^2}\sum_{i\in U}\sum_{j\in U, j\neq i}|y_i||y_j| \\
&\leq \frac{1}{\lambda_1\lambda_2}\frac{N^2}{n_2}\frac{n_2}{N}\frac{1}{N}\sum y_i^2 + \frac{N^2}{n_1n_2}\frac{C_1C_2}{\lambda_1^2\lambda_2^2}\times\left(\frac{1}{N}\sum_{i\in U}|y_i|\right)^2 \\
&\leq C\frac{N^2}{n_2} + C^*\frac{N^2}{n_1n_2}.
\end{aligned}$$

We conclude that $|A_3| = O\left(\frac{N^2}{n_2}\right)$ and $|A_2| = O\left(\frac{N^2}{n_2}\right)$. It follows that $V_p(\hat{Y}_{DE}) = O\left(\frac{N^2}{n_2}\right)$ and so

$$\frac{1}{N}\left(\hat{Y}_{DE} - Y\right) = O_p\left(n_2^{-1/2}\right). \tag{A.1}$$

We now study the term $\frac{1}{2}(\hat{B}^{DE}_{min} + \hat{B}^{DE}_{max})$ on the right hand side of (4.4.5).

$$\begin{aligned}
|\hat{B}_i^{DE}| &= |\sum_{j\in S_2} \frac{\pi_i^*}{\pi_{ij}^*}\left(\frac{\pi_{ij}^*}{\pi_i^*\pi_j^*} - 1\right) y_j| \\
&= |\sum_{j\in S_2} \frac{\pi_{1i}\pi_{2i}}{\pi_{1ij}\pi_{2ij}}\left(\frac{\pi_{1ij}\pi_{2ij} - \pi_{1i}\pi_{2i}\pi_{1j}\pi_{2j}}{\pi_{1i}\pi_{2i}\pi_{1j}\pi_{2j}}\right) y_j| \\
&\leq \frac{1-\pi_{1i}\pi_{2i}}{\pi_{1i}\pi_{2i}}|y_i| + \frac{1}{\lambda_1^*\lambda_2^*\lambda_1^2\lambda_2^2} \sum_{j\in S_2, j\neq i} |\left(\pi_{1ij}\pi_{2ij} - \pi_{1i}\pi_{2i}\pi_{1j}\pi_{2j}\right)||y_j|.
\end{aligned}$$

Since

$$\begin{aligned}
\left(\pi_{1ij}\pi_{2ij} - \pi_{1i}\pi_{2i}\pi_{1j}\pi_{2j}\right) &= \left(\pi_{1ij} - \pi_{1i}\pi_{1j}\right)\left(\pi_{2ij} - \pi_{2i}\pi_{2j}\right) \\
&+ \pi_{2i}\pi_{2j}\left(\pi_{1ij} - \pi_{1i}\pi_{1j}\right) + \pi_{1i}\pi_{1j}\left(\pi_{2ij} - \pi_{2i}\pi_{2j}\right),
\end{aligned}$$

we have

$$\begin{aligned}
\max_{i\neq j}|\left(\pi_{1ij}\pi_{2ij} - \pi_{1i}\pi_{2i}\pi_{1j}\pi_{2j}\right)| &\leq \max_{i\neq j}|\left(\pi_{1ij} - \pi_{1i}\pi_{1j}\right)\left(\pi_{2ij} - \pi_{2i}\pi_{2j}\right)| \\
&+ \max_{i\neq j}|\pi_{2i}\pi_{2j}\left(\pi_{1ij} - \pi_{1i}\pi_{1j}\right)| \\
&+ \max_{i\neq j}|\pi_{1i}\pi_{1j}\left(\pi_{2ij} - \pi_{2i}\pi_{2j}\right)| \\
&\leq \frac{C_1C_2}{n_1n_2} + \frac{C_1}{n_1} + \frac{C_2}{n_2} \\
&\leq \frac{C}{n_2}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\frac{1}{N}|\hat{B}_i^{DE}| &\leq \tfrac{1}{N}|\ \frac{1-\pi_{1i}\pi_{2i}}{\pi_{1i}\pi_{2i}}|y_i| + \frac{1}{\lambda_1^*\lambda_2^*\lambda_1^2\lambda_2^2}\frac{C}{n_2}\frac{1}{N}\sum_{j\in S_2, j\neq i}|y_j| \\
&\leq \frac{1}{N}\frac{1-\lambda_1\lambda_2}{\pi_{1i}\pi_{2i}}|y_i| + \frac{1}{\lambda_1^*\lambda_2^*\lambda_1^2\lambda_2^2}\frac{C}{n_2}\frac{1}{N}\sum_{j\in U}|y_j| \\
&\leq \frac{1}{N}\frac{1-\lambda_1\lambda_2}{\pi_{1i}\pi_{2i}}|y_i| + \frac{1}{\lambda_1^*\lambda_2^*\lambda_1^2\lambda_2^2}\frac{C'}{n_2}.
\end{aligned}$$

Assuming that $\max_{i\in U}\frac{1}{\pi_{1i}\pi_{2i}} = O(\frac{N}{n_2})$ and $\max_{i\in U}(y_i) = O(N^\rho)$ with $0 \le \rho \le \frac{1}{2}$, we obtain

$$\begin{aligned}
\frac{1}{N}\max_{i\in S_2}|\hat{B}_i^{DE}| &\le (1-\lambda_1\lambda_2)\frac{C''}{n_2}\max_{i\in U}(|y_i|) + \frac{1}{\lambda_1^*\lambda_2^*\lambda_1^2\lambda_2^2}\frac{C'}{n_2} \\
&\le (1-\lambda_1\lambda_2)C'''\frac{1}{n_2}O(N^\rho) + \frac{1}{\lambda_1^*\lambda_2^*\lambda_1^2\lambda_2^2}\frac{C'}{n_2} \\
&\le (1-\lambda_1\lambda_2)C''''\frac{1}{n_2}O(n_2^\rho) + \frac{1}{\lambda_1^*\lambda_2^*\lambda_1^2\lambda_2^2}\frac{C'}{n_2}.
\end{aligned}$$

Therefore,

$$|\frac{1}{2N}(\hat{B}_{min}^{DE} + \hat{B}_{max}^{DE})| = O\left(n_2^{\rho-1}\right). \tag{A.2}$$

Combining (A.1) and (A.2), we obtain

$$\frac{1}{N}\left(\hat{Y}_{DE}^R(c_{opt}) - Y\right) = O_p\left(n_2^{-1/2}\right)$$

References

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. To appear in *Biometrika*.

Beaumont, J.-F. and Rivest, L.-P. (2009). Dealing with outliers with survey data. Handbook of Statistics, Volume 29, Chapter 11, Sample Surveys: Theory Methods and Inference, Editors: C.R. Rao and D. Pfeffermann, 247–279.

Chambers, R. L., Kokic, P., Smith, P. and Cruddas, M. (2000). Winsorization for identifying and treating outliers in business surveys. *Proc. of the Second International Conference on Establishment Surveys, Am. Stat. Assoc., Alexandria, Virginia*, 717–726.

Elliott, M.R. and R.J.A., Little (2000). Model-Based Alternatives to Trimming Survey Weights. *Journal of Official Statistics*, 16, 191–209.

Estevao, V. M., and Särndal, C. E. (2002). Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling. *Journal of Official Statistics*, 18, 233–255.

Esteavo, V.M. and Särndal, C.-E. (2006). Survey Estimates by Calibration on Complex Auxiliary Information. *International Statistical Review*, 74, 127–147.

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79–87.

Kim, J.K. and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 501–514.

Kokic, P.N., and Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, 10, 419–435.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M. and Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: conditional bias. *Biometrika*, 86, 923–968.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M., Jimenez-Gamero, M.D. and Muñoz-Pichardo, J. (2002). Influence diagnostics in survey sampling: estimating the conditional bias. *Metrika*, 55, 209–214.

Rivest, L.-P. and Hurtubise, D. (1995). On Searls Winsorized means for skewed populations. *Survey Methodology*, 21, 119–129.

Zaslavsky, A.M., Schenker, N. and Belin, T.R. (2001). Downweighting influential clusters in surveys: application to the 1990 post enumeration survey. *Journal of the American Statistical Association*, 96, 858–869.

You Y., Rao J.N.K. and Dick P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631–640.

Chapter 5

Robust estimation for infinite skewed populations: a general approach

Résumé

L'estimation de la moyenne dans le cas d'une population asymétrique est un problème très important en pratique. En effet, il est très courant d'observer des variables dont la distribution est asymétrique, c'est le cas par exemple du chiffre d'affaire des entreprises ou le revenu des ménages. En pratique, l'échantillon d'observation possède des unités qui sont très influentes sur la moyenne empirique, qui est l'estimateur souvent privilégié. Rivest (1994) propose un estimateur non paramétrique pour la moyenne d'une population asymétrique en winsorisant la plus grande ou les deux plus grandes observations de l'échantillon. Il montre que cet estimateur possède de bonnes propriétés en termes d'erreur quadratique moyenne. Sa stratégie consiste à réduire voire supprimer l'influence des plus grandes valeurs de l'échantillon. Notre démarche consiste à quantifier l'influence des unités de l'échantillon et de construire un estimateur robuste en réduisant l'impact des unités influentes. Pour cela, nous allons utiliser le biais conditionnel comme mesure d'influence. Nous donnerons les propriétés de cet estimateur en terme d'erreur quadratique moyenne, nous développerons une approximation de cette erreur quadratique moyenne suivant les différents domaines d'attraction possibles pour la loi considérée et nous effectuerons une étude par simulation pour comparer les performances de cet estimateur avec celui proposé par Rivest (1994).

Mots clés : Biais conditionnel ; domaine d'attraction du maximum ; statistiques d'ordre ; estimation robuste ; loi asymétrique

Abstract

Many variables encountered in practice (e.g., economic variables) have skewed distributions. Estimating the population mean μ of a skewed population is an important problem in practice. While the sample mean is unbiased for the true mean regardless of the underlying distribution that generated the sample observations, it can be highly unstable in the context of skewed distributions. To cope with this problem, we suggest a robust estimator of the population mean based on the concept of conditional bias, which is a measure of influence of a unit. The idea is to reduce the impact of the sample units that have a large influence. The resulting estimator depend on a cut-off value. Following Beaumont et al. (2013), we suggest to select the cut-off value that minimizes the maximum absolute estimated conditional bias with respect to the robust estimator. An estimator of the mean square error of the proposed estimator is also presented. An empirical investigation comparing several estimators in terms of relative bias an relative efficiency shows that the proposed estimator performs well in a variety of scenarios. Finally, the proposed estimator of the mean square error is shown to perform very well empirically in terms of relative bias.

Key Words: conditional bias, max-domain of attraction, order statistics, robust estimation, skewed distribution

5.1 Introduction

Many variables encountered in practice (e.g., economic variables) have skewed distributions. Estimating the mean μ of a skewed population is an important problem in practice. A simple estimator of μ is the sample mean, which is unbiased for μ regardless of the underlying distribution. For this reason, the sample mean can be viewed as a non-parametric estimator of μ. While the sample mean is optimal for the $L2$-norm under some distributions (e.g, the normal and the exponential distributions), it may be highly unstable in the case of highly skewed distributions. Commonly skewed distribution encountered in the literature include the Pareto, the Weibull and the lognormal distributions. Skewness in the sample may also indicate that the latter was generated from a mixture of two distributions (e.g., a mixture of two normal distributions). The lack of stability of the sample mean in the case of highly skewed distributions is due to the possible presence of influential units, which are those units that have a drastic impact on the estimates if they were to be excluded from the sample. When the interest lies in estimating the population mean, the presence of influential units is common when the sample is generated from a highly skewed distribution.

If the underlying distribution that generated the sample was known, one could estimate μ by its maximum likelihood (ML) estimator, which is known to be asymptotically optimal. However, in practice, the underlying distribution is not known. One may attempt to postulate a parametric model and estimate μ by the corresponding ML estimator obtained under the selected model. However, selecting an appropriate model is problematic unless the sample size is large because the influential units are usually found in the tails. If the model assumptions are violated, the ML estimators may be biased and/or inefficient. For example, for the lognormal distribution, Myers and Pepin (1990) showed empirically that the ML estimator of μ was less efficient than the sample mean when the lognormal distri-

bution was slightly misspecified. Thus, it is desirable to develop non-parametric alternatives to the sample mean that are robust to the presence of influential units in the sample. If $\hat{\mu}$ denotes the optimal (ML) estimator of μ with respect to the true model, an estimator is said to be robust when its efficiency is not much affected if the underlying model is slightly misspecified and when its efficiency is close to that of the optimal estimator $\hat{\mu}$ when the model holds.

To guard against the presence of influential units in the sample, Searls (1966) considered the winzorisation technique, which consists of replacing the observations larger than a cut-off value c by c and averaging the resulting observations. Searls (1966) suggested selecting the value of c which minimizes the estimated mean square error of the winsorized estimator. Rivest (1994) suggested a simple alternative that consists of setting c to the second largest observation in the sample. The resulting estimator is called the once-winsorized estimator. Rivest (1994) showed that the once-winzorised estimator is the best possible when the cut-off is selected among extreme order statistics.

In this chapter, we use the conditional bias of a unit as a measure of influence. In the infinite population set-up, the conditional bias was introduced by Muñoz-Pichardo et al. (1995). Beaumont et al. (2013) showed that the conditional bias of a unit is approximately proportional to the customary influence function (Hampel, 1974). The conditional bias can be easily computed for a variety of estimators/parameters, which is an attractive property. In this chapter, we propose an alternative robust estimator based on the concept of conditional bias. The idea is to reduce the impact of the sample units that have a large conditional bias. The proposed estimator depend on a cut-off value. Following Beaumont et al. (2013), we suggest selecting the cut-off value that minimizes the maximum absolute estimated conditional bias with respect to the robust estimator.

2 A robust estimator of the population mean

Let $X_1, ...X_n$ be n independent and identically distributed random variables drawn from a continuous distribution F defined on $\mathbb{R}$. The distribution F is assumed to have a finite mean μ and finite variance σ^2. We denote by $X_{(1)}, .., X_{(n)}$ the order statistics of the sample. Let $\mu_i = \mathbb{E}\left(X_{(i)}\right)$ denote their first moment and $\mu_{i,j} = \mathbb{E}\left(X_{(i)}X_{(j)}\right)$ their moment product. Finally, let $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ denote the sample mean.

Let $\widehat{\mu}$ be an estimator of μ. In order to quantify the influence of unit i on $\widehat{\mu}$, we use the concept of conditional bias introduced by Muñoz-Pichardo et al. (1995).

Definition 5.1. *The conditional bias of unit i on an estimator $\widehat{\mu}$ is defined as*

$$B_i(\widehat{\mu}) = \mathbb{E}(\widehat{\mu}|X_i) - \mu. \tag{5.1.1}$$

The realized value of $B_i(\widehat{\mu})$ conditionally on $X_i = x_i$ will be denoted by $b_i(\widehat{\mu})$.

Proposition 5.1. *When $\widehat{\mu} = \overline{X}$, the conditional bias (5.1.1) reduces to*

$$B_i\left(\overline{X}\right) = \frac{1}{n}\left(X_i - \mu\right). \tag{5.1.2}$$

The proof is given in the Appendix. The realized value of $B_i\left(\overline{X}\right)$ is $b_i\left(\overline{X}\right) = n^{-1}\left(x_i - \mu\right)$, which is unknown as μ is unknown. An estimator of the conditional bias is presented in the next proposition.

Proposition 5.2. *When $\widehat{\mu} = \overline{X}$, a conditionally unbiased estimator of $b_i\left(\overline{X}\right)$ is*

given by

$$\begin{aligned} \widehat{B}_i\left(\overline{X}\right) &= \frac{1}{n}\left(X_i - \overline{X}_{-i}\right) \\ &= \frac{1}{n-1}\left(X_i - \overline{X}\right), \end{aligned} \tag{5.1.3}$$

where

$$\overline{X}_{-i} = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} X_j$$

denotes the sample mean obtained after removing unit i.

The proof is given in the Appendix. It may be tempting to estimate the conditional bias by replacing μ in (5.1.2) by a robust estimator (e.g., an M-estimator). As we argue in Section 5.3 and show empirically in Section 5.5, this has little effect on the properties of the resulting robust estimators.

The error $\overline{X} - \mu$ can be expressed as

$$\overline{X} - \mu = \sum_{j=1}^{n} B_j\left(\overline{X}\right).$$

That is, the conditional bias of a unit can be viewed as its contribution to the error of $\overline{X}$. Also, the variance of $\overline{X}$ can be expressed as

$$\mathbb{V}\left(\overline{X}\right) = \mathbb{E}\left\{\sum_{i=1}^{n} B_i^2\left(\overline{X}\right)\right\}$$

showing the relationship between the conditional bias and the variance of an estimator.

Following Beaumont et al. (2013), we construct a robust version of $\overline{X}$:

$$\overline{X}^R(c) = \overline{X} - \sum_{j=1}^{n} B_j\left(\overline{X}\right) + \sum_{j=1}^{n} \psi_c\left\{B_j\left(\overline{X}\right)\right\}, \tag{5.1.4}$$

where ψ_c denotes the Huber function defined as $\psi_c(x) = \min\{|c|, \max(-|c|, x)\}$ and c is a tuning constant whose value must be determined. One option is to determine the value of c that minimizes the estimated mean square error of the robust estimator $\overline{X}^R(c)$. Generally, this involves messy calculations and is often achieved at the expense of simplifying assumptions. To cope with the problem, we consider an alternative criterion which consists on finding the value of c that minimizes the maximum absolute estimated conditional bias with respect of the robust estimator $\overline{X}^R(c)$. More specifically, the optimal value of c, denoted by c_{opt}, is given by

$$c_{opt} \quad = \quad \arg\min_{c\in\mathbb{R}} \left(\max\left[\left|\widehat{B}_i\left\{\overline{X}^R(c)\right\}\right|, i \in [1,n]\right]\right),$$

where $\widehat{B}_i\left\{\overline{X}^R(c)\right\}$ is an estimator of $b_i\left\{\overline{X}^R(c)\right\}$ defined as

$$\begin{aligned} b_i\left\{\overline{X}^R(c)\right\} =& \mathbb{E}\left\{\overline{X}^R(c)\,|X_i = x_i\right\} - \mu \\ =& \mathbb{E}\left(\overline{X}|X_i = x_i\right) + \mathbb{E}\left(\sum_{j=1}^{n}\left[\psi_c\left\{B_j\left(\overline{X}\right)\right\} - B_j\left(\overline{X}\right)\right]|X_i = x_i\right) - \mu \\ =& b_i\left(\overline{X}\right) + \mathbb{E}\left(\sum_{j=1}^{n}\left[\psi_c\left\{B_j\left(\overline{X}\right)\right\} - B_j\left(\overline{X}\right)\right]|X_i = x_i\right). \end{aligned}$$

Since the latter expectation is unknown, we can estimate it by

$$\widehat{B}_i\left\{\overline{X}^R(c)\right\} = \widehat{B}_i(\overline{X}) + \sum_{j=1}^{n}\left[\psi_c\left(\widehat{B}_j(\overline{X})\right) - \widehat{B}_j(\overline{X})\right], \tag{5.1.5}$$

where $\widehat{B}_i\left(\overline{X}\right)$ is an estimator of $b_i\left(\overline{X}\right)$. If we use the non-robust estimator of the conditional bias given by the expression (5.1.3), the resulting estimator $\widehat{B}_i\left\{\overline{X}^R(c)\right\}$ is conditionally unbiased for $b_i\left\{\overline{X}^R(c)\right\}$ in the sense

$$\mathbb{E}\left[\widehat{B}_i\left\{\overline{X}^R(c)\right\} \mid X_i = x_i\right] = b_i\left\{\overline{X}^R(c)\right\}.$$

The estimated conditional bias (5.1.5) can be written as

$$\widehat{B}_i\left\{\overline{X}^R(c)\right\} = \widehat{B}_i\left(\overline{X}\right) + \Delta(c),$$

where

$$\Delta(c) = \sum_{j=1}^{n}\left[\psi_c\left\{\widehat{B}_j\left(\overline{X}\right)\right\} - \widehat{B}_j\left(\overline{X}\right)\right].$$

Therefore, the value c_{opt} is given by

$$c_{opt} \;=\; \arg\min_{c\in\mathbb{R}}\left[\max\left\{\left|\widehat{B}_i(\overline{X}) + \Delta(c)\right|, i \in [1,n]\right\}\right].$$

It is easily shown that $\Delta(c_{opt})$ is given by

$$\Delta(c_{opt}) = \frac{1}{2}\left[\min\left\{\widehat{B}_i\left(\overline{X}\right)\right\} + \max\left\{\widehat{B}_i\left(\overline{X}\right)\right\}\right].$$

The estimator $\overline{X}^R(c)$ in (5.1.4) evaluated at c_{opt} reduces to

$$\overline{X}^R(c_{opt}) = \overline{X} - \frac{1}{2}\left[\min\left(\widehat{B}_i(\overline{X})\right) + \max\left(\widehat{B}_i(\overline{X})\right)\right]. \tag{5.1.6}$$

Using (5.1.3) in the previous expression, we obtain

$$\overline{X}^R(c_{opt}) = \frac{n}{n-1}\left\{\overline{X} - \frac{1}{n}\left(\frac{X_{(1)} + X_{(n)}}{2}\right)\right\}. \tag{5.1.7}$$

At this stage, it is useful to recall the expression of the once-winzorised estimator of Rivest (1994), which we note $\overline{X}_1^R$:

$$\overline{X}_1^R = \overline{X} - \frac{1}{n}\left(X_{(n)} - X_{(n-1)}\right). \tag{5.1.8}$$

A comparison of (5.1.7) and (5.1.8) shows that both estimators apply different adjustment factors to the sample mean $\overline{X}$. On the one hand, the once-winsorized estimator $\overline{X}_1^R$ subtracts a fraction of the distance between the largest and the second largest observations from the sample mean. On the other hand, ignoring the factor $n/(n-1)$, the proposed robust estimator $\overline{X}^R(c_{opt})$ subtracts a fraction of the midrange $\left(X_{(1)} + X_{(n)}\right)/2$.

5.2 Properties of the robust estimator

In this section, we establish the theoretical properties of the robust estimator (5.1.7).

Proposition 5.3. *The bias of* $\overline{X}^R(c_{opt})$ *is given by*

$$Bias\left\{\overline{X}^R(c_{opt})\right\} = \frac{1}{n-1}\left\{\mu - \left(\frac{\mu_1 + \mu_n}{2}\right)\right\}. \tag{5.2.1}$$

The proof of Proposition 5.3 is trivial and is thus omitted. For symmetric distributions, the bias of $\overline{X}^R(c_{opt})$ vanishes since, in this case, $(\mu_1 + \mu_n)/2 = \mu$. When the distribution is not symmetric, the bias of $\overline{X}^R(c_{opt})$ is virtually untractable. However, it is possible to bound the absolute bias. This is presented in the next proposition.

Proposition 5.4. *Under the chapter assumptions that the distribution* F *is con-*

tinuous with finite mean μ and finite variance σ^2, we have

$$\left|Bias\left\{\overline{X}^R(c_{opt})\right\}\right| \leq \frac{\sigma}{\sqrt{2n-1}}.$$

The proof of Proposition 5.4 is presented in the Appendix.

Proposition 5.5. *The robust estimator $\overline{X}^R(c_{opt})$ is L1-consistent for μ; that is,*

$$\mathbb{E}\left\{\left|\overline{X}^R(c_{opt}) - \mu\right|\right\} \xrightarrow[n\to\infty]{} 0.$$

The proof of Proposition 5.5 is presented in the Appendix.

Theorem 5.1. *The mean square error of $\overline{X}^R(c_{opt})$ is given by*

$$\begin{aligned} MSE\left\{\overline{X}^R(c_{opt})\right\} &= \frac{1}{(n-1)^2}\left\{\left(\mu^2+n\sigma^2\right) - \left(\mu_{n,n} - \mu_{n-1,n} + \mu\mu_{n-1}\right)\right. \\ &+ \left. 4^{-1}\left(\mu_{1,1} + \mu_{n,n} + 2\mu_{1,n}\right) - \left(\mu_{1,1} - \mu_{2,1} + \mu\mu_2\right)\right\} \quad (5.2.2) \end{aligned}$$

Corollary 5.1. *If $X_1, ...X_n$ are drawn from a normal distribution with mean μ and variance σ^2, we have*

$$MSE\left\{\overline{X}^R(c_{opt})\right\} = \frac{n}{(n-1)}\left[\frac{\sigma^2}{n} + \frac{\sigma^2\pi^2}{24n(n-1)log(n)} + O\left\{\frac{1}{n^2log(n)^2}\right\}\right]. \quad (5.2.3)$$

It is interesting to compare the efficiency of the robust estimator $\overline{X}^R(c_{opt})$ to that of the sample mean $\overline{X}$, which is optimal under the normal distribution. Using

(5.2.3), the relative efficiency of $\overline{X}^R(c_{opt})$ *is given by*

$$RE\left\{\overline{X}^R(c_{opt})\right\} = \frac{MSE\left\{\overline{X}^R(c_{opt})\right\}}{MSE\left(\overline{X}\right)} = \frac{n}{(n-1)}\left[1 + \frac{\pi^2}{24(n-1)log(n)} + O\left\{\frac{1}{nlog(n)^2}\right\}\right].$$

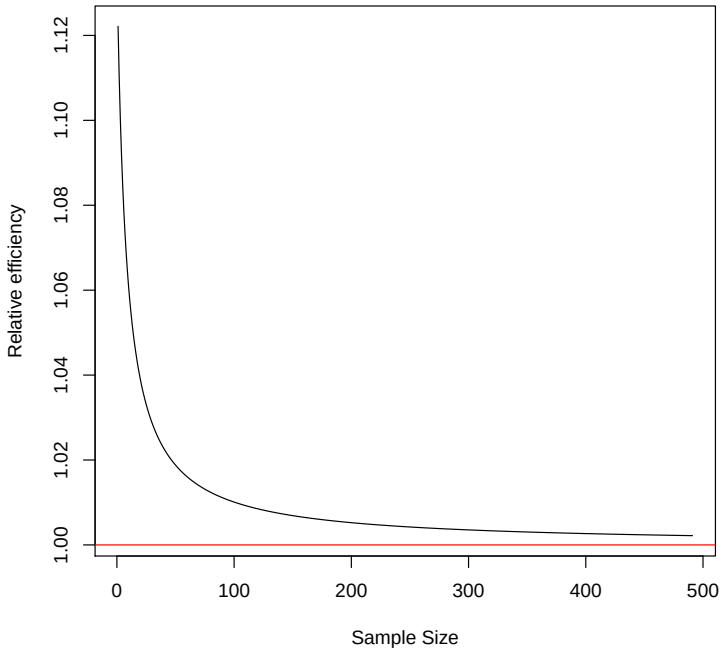


Figure 5.2.1: Relative efficiency of $\overline{X}^R(c_{opt})$ as a function of n

It is clear from Figure 5.2.1 that for $n \geq 50$, the loss of efficiency of $\overline{X}^R(c_{opt})$ is very small (less than 2%). Next, we give the expression of the mean square error of $\overline{X}^R(c_{opt})$ for distributions (e.g. the Weibull, the gamma and the lognormal distributions) belonging to the max-domain of attraction Gumbel.

Corollary 5.2. *We assume that* F *belongs to the max-domain of attraction Gumbel. In addition, we make the following assumptions:*

(i) $supp(F) = [w_{min}, +\infty[$, *where* $-\infty < w_{min}$;

(ii) There exists $(a,\ K,\ alpha) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{+*}$ *such that the density function* f *has the form* $f(x) = K\left(x - w_{min}\right)^{\alpha}$ *on* $[w_{min}; a]$.

In this case, the mean square error of $\overline{X}^{R}\left(c_{opt}\right)$ *given by (5.2.2) becomes*

$$\begin{aligned} MSE\left\{\overline{X}^{R}\left(c_{opt}\right)\right\} =& \frac{1}{(n-1)^2}\left\{n\sigma^2 - h^2\left\{log(n)\right\}\left(g_1\left\{log(n)\right\} + 2\left[g_1^2\left\{log(n)\right\} + g_2\left\{log(n)\right\}\right]\right.\right. \\ &+ 12g_1\left\{log(n)\right\}g_2\left\{log(n)\right\}) - \mu h\left\{log(n)\right\}\Big\} \\ &+ \frac{1}{4(n-1)^2}\left(h^2\left\{log(n)\right\}\left[1 + 2g_1\left\{log(n)\right\} + 2g_1^2\left\{log(n)\right\}\right.\right. \\ &+ 12g_1\left\{log(n)\right\}g_2\left\{log(n)\right\} + 4g_2\left\{log(n)\right\}] \\ &+ 2w_{min}h\left\{log(n)\right\}\left[1 + g_1\left\{log(n)\right\} + 2g_2\left\{log(n)\right\}\right]) \\ &+ O\left[\frac{h^2\left\{log(n)\right\}g_1^3\left\{log(n)\right\}}{n^2}\right], \end{aligned} \tag{5.2.4}$$

where

$$g_1(t) = \left[e^t f\left\{F^{-1}\left(1 - e^{-t}\right)\right\}F^{-1}\left(1 - e^{-t}\right)\right],$$

$$g_2\left(t\right) = -\frac{1}{2}\left[g_1\left(t\right) + g_1^2\left(t\right)\frac{f'\left\{F^{-1}\left(1 - e^{-t}\right)\right\}F^{-1}\left(1 - e^{-t}\right)}{f\left\{F^{-1}\left(1 - e^{-t}\right)\right\}}\right]$$

and

$$h\left(t\right) = F^{-1}\left(1 - e^{-t}\right).$$

We consider the special case of the Weibull distribution with shape parameter k and scale parameter λ, which belongs to the max-domain of attraction of Gumbel. The distribution function is given by

$$F(x) = 1 - exp\left\{-\left(\frac{x}{\lambda}\right)^k\right\}.$$

In this case, we have $h(t) = \lambda t^{\frac{1}{k}}$, $g_1(t) = \frac{1}{kt}$ and $g_2(t) = \frac{1-k}{2k^2t^2}$. Ignoring the terms of order $O\left(\frac{log(n)^{\frac{2}{k}-3}}{n^2}\right)$, the expression (5.2.4) reduces to

$$\begin{aligned} MSE(\overline{X}^R(c_{opt}) &= \frac{n^2}{(n-1)^2}\left[\frac{\sigma^2}{n} - \frac{\lambda^2 log(n)^{\frac{2}{k}}}{n^2}\left\{-\frac{1}{4} + \frac{1}{2klog(n)}\right.\right. \\ &+ \left.\left.\left(2-\frac{k}{2}\right)\frac{1}{k^2log(n)^2} + \frac{\Gamma(1+\frac{1}{k})}{log(n)^{\frac{1}{k}}}\right\}\right]. \end{aligned} \tag{5.2.5}$$

It is interesting to compare (5.2.5) with the $O\left(\frac{log(n)^{\frac{2}{k}-3}}{n^2}\right)$ mean square error approximation of the once-winzorised estimator of Rivest (1994, p.377):

$$MSE(\overline{X_1}^R) = \frac{\sigma^2}{n} - \frac{2\lambda^2(1-k)log(n)^{\frac{2}{k}-2}}{k^2n^2}.$$

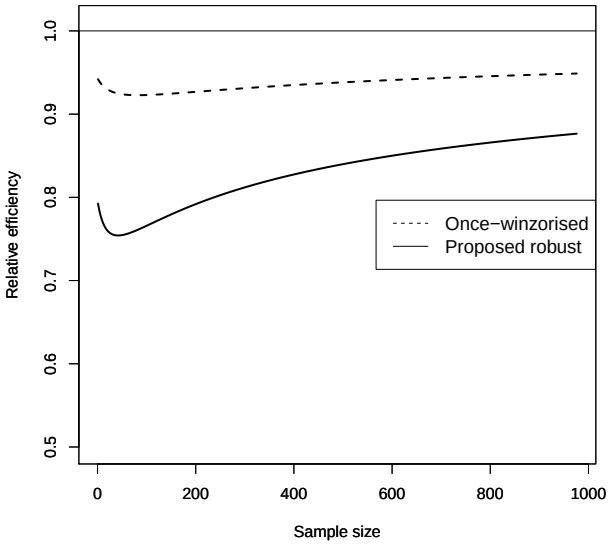
Using the sample mean as the reference, the relative efficiency of $\overline{X}^R(c_{opt})$ and $\overline{X_1}^R$ are respectively given by

$$\begin{aligned} RE\left\{\overline{X}^R(c_{opt})\right\} = \frac{MSE\left\{\overline{X}^R(c_{opt})\right\}}{MSE\left(\overline{X}\right)} &= 1 - \frac{log(n)^{\frac{2}{k}}}{n\left\{\Gamma\left(1+\frac{2}{k}\right) - \Gamma\left(1+\frac{1}{k}\right)^2\right\}}\left\{-\frac{1}{4} + \frac{1}{2klog(n)}\right. \\ &\left. + \left(2-\frac{k}{2}\right)\frac{1}{k^2log(n)^2} + \frac{\Gamma(1+\frac{1}{k})}{log(n)^{\frac{1}{k}}}\right\} \end{aligned}$$

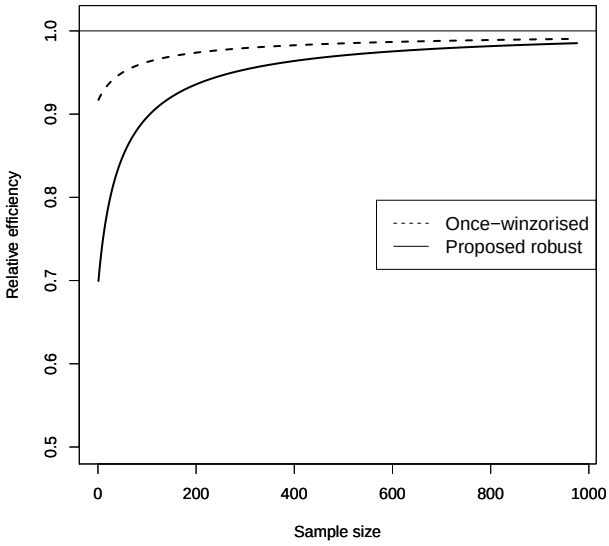
and

$$RE\left(\overline{X_1}^R\right) = \frac{MSE\left(\overline{X_1}^R\right)}{MSE\left(\overline{X}\right)} = 1 - \frac{2(1-k)log(n)^{\frac{2}{k}-2}}{k^2n\left\{\Gamma\left(1+\frac{2}{k}\right) - \Gamma\left(1+\frac{1}{k}\right)^2\right\}}.$$

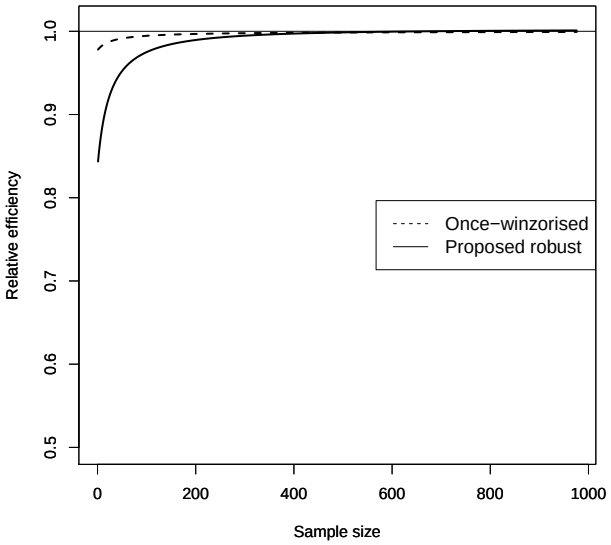
Note that the previous two expressions do not depend on the parameters λ and σ^2, due to the fact that $\lambda^2/\sigma^2 = \Gamma\left(1+\frac{2}{k}\right) - \Gamma\left(1+\frac{1}{k}\right)^2$.



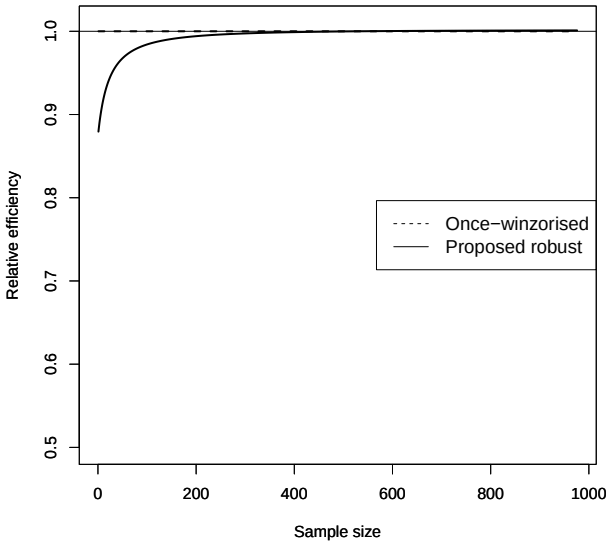
(a) $k = 0.3$



(b) $k = 0.5$



(c) $k = 0.8$



(d) $k = 1$, exponential distribution

Figure 5.2.2: Relative efficiency of $\overline{X}^R(c_{opt})$ as a function of n for the Weibull distribution

Figure 5.2.2 shows the relative efficiency of the Once-winsorized and the proposed robust estimator for values of n ranging from 25 to 1000 and for different values of the Weibull shape parameter k. Note that for k equal to 1, the Weibull distribution is equivalent to the Exponential distribution. As we can see in Figure 5.2.2, the gain in efficiency of both estimators increase with $1/k$, thus an increase in skewness leads to a higher gain in efficiency. We also notice that the relative efficiency tends to 1, since they are both asymptotically equivalent to the sample mean. Finally, Figure 5.2.2 shows that the proposed robust estimator is more efficient than the Once-winsorized estimator for any value of k.

Finally, we give the expression of the mean square error of $\overline{X}^R(c_{opt})$ for the Pareto distribution that belongs to the max-domain of attraction Frechet.

Corollary 5.3. *Assuming that $X_1, ...X_n$ are drawn from a Pareto distribution with shape parameter γ and scale parameter 1. In this case, the mean square error of $\overline{X}^R(c_{opt})$ given by (5.2.2) reduces to*

$$MSE\left\{\overline{X}^R(c_{opt})\right\} = \frac{n\sigma^2}{(n-1)^2} + \frac{1}{(n-1)^2}\left(\frac{1}{4} - \frac{1}{\gamma - 1}\right)\mu_{n,n} + \frac{1}{(n-1)^2}\left\{\frac{\mu_{1,n}}{2} - \mu\left(1 - \frac{1}{\gamma}\right)\mu_n\right\}. \tag{5.2.6}$$

Applying the Sterling formula to the Gamma function, we have

$$\mu_n \simeq n^{1/\gamma}\Gamma\left(1 - 1/\gamma\right),$$

$$\mu_{n,n} \simeq n^{2/\gamma}\Gamma\left(1 - 2/\gamma\right)$$

and

$$\mu_{1,n} = n^{1/\gamma}\Gamma\left(1 - 1/\gamma\right)$$

Using these approximations, the mean square error (5.2.6) reduces to

$$MSE\left\{\overline{X}^{R}\left(c_{opt}\right)\right\} \simeq \frac{n\sigma^{2}}{(n-1)^{2}}+\frac{n^{2/\gamma}\Gamma\left(1-2/\gamma\right)}{(n-1)^{2}}\left(\frac{1}{4}-\frac{1}{\gamma-1}\right)+\frac{n^{1/\gamma}\Gamma\left(1-1/\gamma\right)}{(n-1)^{2}}\left\{\frac{1}{2}-\mu\left(1-\frac{1}{\gamma}\right)\right\}. \tag{5.2.7}$$

It is interesting to compare (5.2.7) with the mean square error approximation of the once-winzorised estimator of Rivest (1994, p.377):

$$MSE(\overline{X_{1}}^{R}) \simeq \frac{\sigma^{2}}{n}-\frac{2\Gamma\left(1-2/\gamma\right)}{n^{2-2/\gamma}\gamma\left(\gamma-1\right)}.$$

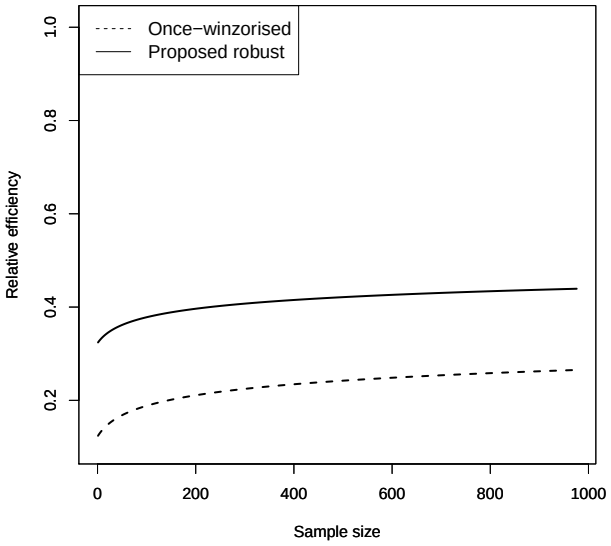
Therefore, the relative efficiency of $\overline{X}^{R}\left(c_{opt}\right)$ can be approximated by

$$\begin{aligned} RE\left\{\overline{X}^{R}\left(c_{opt}\right)\right\} &= \frac{MSE\left\{\overline{X}^{R}\left(c_{opt}\right)\right\}}{MSE\left(\overline{X}\right)} \simeq 1-\frac{n^{2/\gamma-1}\Gamma\left(1-2/\gamma\right)}{\sigma^{2}}\left(\frac{1}{4}-\frac{1}{\gamma-1}\right) \\ &+ \frac{n^{1/\gamma-1}\Gamma\left(1-1/\gamma\right)}{\sigma^{2}}\left\{\frac{1}{2}-\mu\left(1-\frac{1}{\gamma}\right)\right\}, \end{aligned}$$

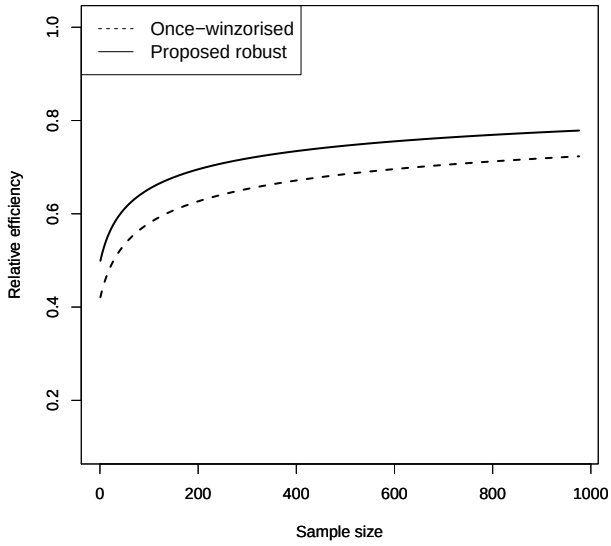
and

$$RE\left(\overline{X_{1}}^{R}\right)=\frac{MSE\left(\overline{X_{1}}^{R}\right)}{MSE\left(\overline{X}\right)}=1-\frac{2\Gamma\left(1-2/\gamma\right)}{\gamma\left(\gamma-1\right)\sigma^{2}n^{1-2/\gamma}}.$$

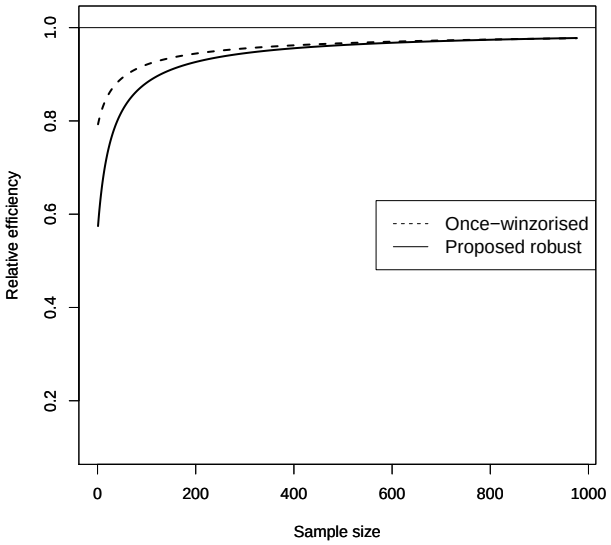
Figure 5.2.3 shows the relative efficiency of the Once-winsorized and the proposed robust estimator for values of n ranging from 25 to 1000 and for different values of the Pareto shape parameter γ. As we can see in Figure 5.2.3, the gain in efficiency of both estimators increase with $1/\gamma$, thus an increase in skewness leads to a higher gain in efficiency. We also notice that the relative efficiency tends to 1, since they are both asymptotically equivalent to the sample mean. Finally, Figure 5.2.3 shows that the Once-winsorized estimator is more efficient than the proposed robust estimator for any value of k.



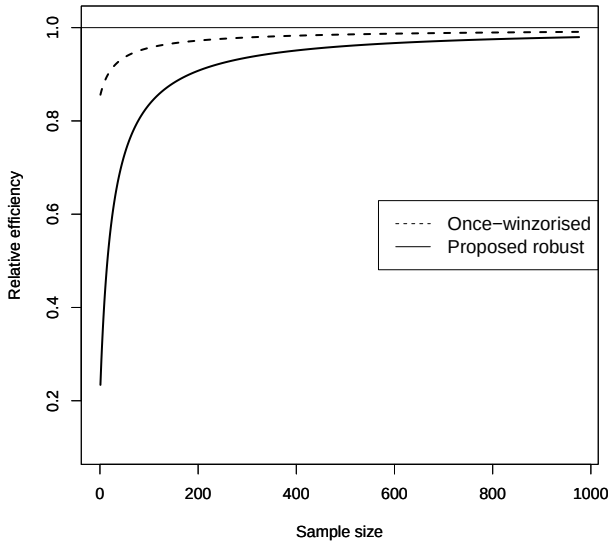
(a) $\gamma = 2.1$



(b) $\gamma = 2.5$



(c) $\gamma = 5$



(d) $\gamma = 8$

Figure 5.2.3: Relative efficiency of $\overline{X}^R(c_{opt})$ as a function of n for the Pareto distribution

Remark 5.1. The estimator (5.1.7) was obtained by estimating the conditional bias in a nonrobust fashion. Instead of replacing μ in (5.1.2) by $\overline{X}_{-i}$ (see Proposition 3), suppose that we use a robust estimator of μ, say an M-estimator. Let $\overline{X}^{R}(c_{opt})$ and $\overline{X}^{R,M}(c_{opt})$ denote respectively the estimator (5.1.6) in which the conditional bias is estimated in a nonrobust fashion and in a robust fashion. Then we have

$$\overline{X}^{R}(c_{opt}) - \overline{X}^{R,M}(c_{opt}) = \frac{1}{n}\left\{\left(\overline{X} - \mu\right) - \left(\hat{\mu}^{R,M} - \mu\right)\right\},$$

where $\hat{\mu}^{R,M}$ denotes an M-estimator of μ. Since M-estimators have order $n^{-1/2}$ consistency (Maronna et al., 2006, p. 45), we have $\overline{X}^{R} - \overline{X}^{R,M} = O_p\left(n^{-3/2}\right)$. Thus, the difference between $\overline{X}^{R}(c_{opt})$ and $\overline{X}^{R,M}(c_{opt})$ is negligible for large n compared to the rate of convergence of the estimator $\overline{X}^{R}(c_{opt})$, since $\overline{X}^{R}(c_{opt}) - \mu = O_p\left(n^{-1/2}\right)$.

5.3 Mean square Error estimation

Since the order statistics appearing in the expression 5.2.2 of the mean square error of the proposed estimator are unknown, it might be useful to propose an estimation of this mean square error. We construct a mean squared error estimator by replacing the expectation of the moment of the order statistics by the observed order statistic in the general expression (5.2.2) of the mean square error. We

obtain

$$\widehat{MSE}\left\{\overline{X}^{R}\left(c_{opt}\right)\right\} = \frac{1}{(n-1)^2}\left\{\left(\overline{X}^{R}\right)^2 + ns^2\right\} - \frac{1}{(n-1)^2}\left\{X_{(n)}\left(X_{(n)} - X_{(n-1)}\right) + \overline{X}^{R}X_{(n-1)}\right\} + \frac{1}{4\left(n-1\right)^2}\left(X_{(1)}^2 + X_{(n)}^2 + 2X_{(1)}X_{(n)}\right) - \frac{1}{(n-1)^2}\left\{X_{(1)}\left(X_{(1)} - X_{(2)}\right) + \overline{X}^{R}X_{(2)}\right\}, \tag{5.3.1}$$

where $s^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. In Section 5.4.2, we assess the performance of (5.3.1) in terms of relative bias.

5.4 Simulation study

5.4.1 Point estimation

We carried out a simulation study to examine the properties of several robust estimators. We generated samples from eight distinct distributions. In each case $J = 100,000$ samples, of size $n = 25; 50; 100; 1000$ were generated. The sampled x-values were generated according to the following model:

$$X_i = U_i + \delta_i V_i,$$

where U_i, δ_i and V_i are random variables, whose distributions are described in Table 5.4.1. The parameters were chosen so that $\mu = 100$.

We were interested in estimating the population mean μ. In each sample, we computed seven estimators: (1) the sample mean $\overline{X}$; (2) the ML estimator obtained for each distribution $\overline{X}_{opt}$; (3) the once-winzorised estimator $\overline{X}_1^R$ given by (5.1.8); (4) the proposed robust estimator $\overline{X}^{R}\left(c_{opt}\right)$ given by (5.1.6), where the conditional bias $B_i(\overline{X}) = \frac{1}{n}\left(X_i - \mu\right)$ given by (5.1.2) was estimated with μ

replaced by (i) the sample mean $\overline{X}$; (ii) the sample median; (iii) an M-estimator where the tuning constant was set to 1.2 and (iv) an M-estimator where the tuning constant was set to 1.5.

Let $\widehat{\mu}$ be a generic estimator of μ. As a measure of bias, we computed the Monte Carlo percent relative bias given by

$$RB_{MC}\left(\widehat{\mu}\right) = \frac{E_{MC}\left(\widehat{\mu}\right) - \mu}{\mu} \times 100,$$

where

$$E_{MC}\left(\widehat{\mu}\right) = \frac{1}{J}\sum_{j=1}^{J}\widehat{\mu}^{(j)}$$

with $\widehat{\mu}^{(j)}$ denoting the estimator $\widehat{\mu}$ in the j-th iteration, $j = 1, \ldots, 100,000$. As a measure of efficiency, using the sample mean $\overline{X}$ as the reference, we computed

$$RE_{MC}\left(\widehat{\mu}, \overline{X}\right) = \frac{\frac{1}{J}\sum_{j=1}^{J}\left(\widehat{\mu}^{(j)} - \mu\right)^2}{\frac{1}{J}\sum_{j=1}^{J}\left(\overline{X}^{(j)} - \mu\right)^2} \times 100,$$

where $\overline{X}^{(j)}$ denotes the sample mean in the j-th iteration, $j = 1, \ldots, 100,000$.

Population	U_i distribution	Mixture	δ_i distribution	V_i distribution
1	$\mathcal{N}(100,\ 2500)$	No	0	0
2	$\mathcal{E}xp(1/100)$	No	0	0
3	$\mathcal{L}$og-$\mathcal{N}\left\{log(100) - (1.2)^2, 1.2\right\}$	No	0	0
4	$\mathcal{L}aplace(100, 5)$	No	0	0
5	$\mathcal{W}eibull(0.5, 50)$	No	0	0
6	$\mathcal{P}areto(60, 2.5)$	No	0	0
7	$\mathcal{N}(50,\ 2500)$	Yes	$\mathcal{B}(0.05)$	$\mathcal{N}(1050,\ 2500)$
8	$\mathcal{N}(52.6,\ 2500)$	Yes	$\mathcal{B}(0.05)$	$\mathcal{L}$og-$\mathcal{N}(6.2, 1.2)$

Table 5.4.1: Models used to generate the populations

Table 5.4.2 shows the relative bias and relative efficiency of the estimators described above. We start by noting that the relative efficiency of both the once-winzorised estimator and the proposed estimator decreased as the sample size n increased. The relative efficiency of both estimators approached 100 as the sample size increased. Also, it is interesting to note that the way of estimating the conditional bias $B_i(\overline{X}) = \frac{1}{n}(X_i - \mu)$ in (5.1.6) had very little effect on both the relative bias and relative efficiency in all the scenarios, which is consistent with remark 5.1. Therefore, from this point on, we focus on the proposed estimator for which the conditional bias was estimated in a nonrobust fashion.

For the normal population, where the sample mean is also the ML estimator, we note that both the once-winzorised estimator and the proposed estimator showed negligible bias and were nearly as efficient as the sample mean. This was true for all the sample sizes.

For the exponential distribution, for which the sample mean was, once again, the ML estimator, both the once-winzorised estimator and the proposed estimator showed some bias. For example, the once-winzorised estimator showed a relative bias equal to -2.1% for $n = 50$, whereas the proposed estimator showed a relative bias equal to -2.7%. Also, both estimators were never less efficient than the sample mean for all the sample sizes.

For the lognormal distribution, both the once-winzorised estimator and the proposed estimator showed very similar properties in terms of relative bias and relative efficiency. For small sample sizes, both estimators were more efficient than the sample mean and were even more efficient than the ML estimator. This can be explained because ML estimators are only asymptotically optimal, therefore requiring a large sample size. For example $n = 1000$, the ML estimator was more efficient (with a relative efficiency equal to 76%) than both the once-winzorised

and the proposed estimators (with a relative efficiency equal to 95%).

Similar remarks can be made for the Weibull and Pareto distributions except that for the Weibull distribution, the proposed estimator slightly outperformed the once-winzorised estimator in terms of relative efficiency, whereas we observed the opposite for the Pareto distribution. This is consistent with the results presented in Figures 5.2.2 and 5.2.3.

For the double exponential distribution, we note that, although both the once-winzorised and the proposed estimators were never less efficient than the sample mean with values of relative efficiency ranging from 92% to 100%, they were less efficient than the ML estimator in all the scenarios.

For the mixture distribution 1, we note that the once-winzorised estimator was less efficient than the sample mean with values of relative efficiency ranging from 100% to 133%. In contrast, for the proposed estimator, the values of relative efficiency were ranging from 88% to 102%.

Finally, for the mixture distribution 2, both the once-winzorised and the proposed estimator were more efficient than the sample mean in all the scenarios. For $n = 25$, the once-winzorised estimator was more efficient than the proposed estimator, although it was significantly more biased. The differences between both estimators vanished as the sample size increased.

Distribution	Sample size	$\overline{X}_{Opt}$	$\overline{X}_1^R$	$\overline{X}^R(c_{opt})$			
				sample mean	median	M-estimator (1.2)	M-estimator (1.5)
Normal	25	−0.1(100)	−0.1(101)	−0.1(100)	−0.1(100)	−0.1(100)	−0.1(100)
	50	−0.0(100)	−0.4(100)	−0.0(100)	−0.0(100)	−0.0(100)	−0.0(100)
	100	0.0(100)	−0.1(100)	0.0(100)	0.0(100)	0.0(100)	0.0(100)
	1000	0.0(100)	−0.0(100)	0.0(100)	0.0(100)	0.0(100)	0.0(100)
Exponential	25	−0.1(100)	−4.1(100)	−4.0(98)	−5.0(98)	−4.5(98)	−4.4(98)
	50	−0.2(100)	−2.1(100)	−2.7(100)	−3.3(100)	−3.1(100)	−3.0(100)
	100	−0.01(100)	−1.1(100)	−1.7(100)	−2.0(100)	−1.8(100)	−1.8(100)
	1000	0.0(100)	−0.1(100)	−0.2(100)	−0.3(100)	−0.3(100)	−0.3(100)
Lognormal	25	2.0(90)	−11.7(74)	−9.2(73)	−10.8(72)	−10.4(72)	−10.3(72)
	50	1.4(81)	−7.2(74)	−6.5(74)	−7.4(74)	−7.2(74)	−7.1(74)
	100	0.7(83)	−4.4(85)	−4.6(84)	−5.1(85)	−4.1(84)	−5.0(84)
	1000	−0.0(76)	−0.9(95)	−1.2(95)	−1.3(95)	−1.3(95)	−1.3(95)
Double Exponential	25	−0.0(67)	−0.2(100)	−0.0(92)	−0.0(90)	−0.0(91)	−0.0(91)
	50	−0.0(61)	−0.1(100)	−0.0(94)	−0.0(94)	−0.0(94)	−0.0(94)
	100	−0.0(60)	−0.0(100)	−0.0(96)	−0.0(96)	−0.0(96)	−0.0(96)
	1000	0.0(54)	−0.0(99)	−0.0(99)	−0.0(99)	−0.0(99)	−0.0(99)
Weibull	25	0.9(100)	−15.3(81)	−12.7(77)	−17.8(76)	−17.8(76)	−17.8(76)
	50	0.6(96)i	−8.9(87)	−9.0(84)	−10.6(84)	−10.6(84)	−10.6(84)
	100	0.4(95)	−5.1(92)	−6.1(89)	−6.8(90)	−6.8(90)	−6.8(90)
	1000	0.1(93)	−0.7(98)	−1.3(98)	−0.4(100)	−0.4(100)	−0.4(100)
Pareto	25	0.7(76)	−5.3(54)	−3.9(61)	−4.6(60)	−4.6(60)	−4.6(60)
	50	0.4(65)	−3.3(57)	−2.9(63)	−3.2(63)	−3.2(63)	−3.2(63)
	100	0.0(61)	−2.3(62)	−2.2(68)	−2.3(68)	−2.3(68)	−2.3(68)
	1000	0.0(59)	−0.6(77)	−0.6(80)	−0.7(81)	−0.7(81)	−0.7(81)
Mixture 1	25	−0.4(100)	−15.3(133)	−12.4(88)	−13.8(86)	−13.7(86)	−13.7(86)
	50	−0.3(100)	−5.2(126)	−8.2(100)	−9.0(99)	−9.0(99)	−9.0(99)
	100	−0.2(100)	−1.2(106)	−4.6(102)	−5.0(103)	−5.0(103)	−5.0(103)
	1000	−0.0(100)	−0.1(100)	−0.6(100)	−0.6(100)	−0.6(100)	−0.6(100)
Mixture 2	25	−1.1(100)	−31.8(32)	−17.7(42)	−18.8(40)	−18.8(40)	−18.8(40)
	50	−0.6(100)	−24.1(44)	−15.5(48)	−16.0(47)	−16.0(47)	−16.0(47)
	100	−1.3(100)	−17.2(57)	−13.2(57)	−13.5(57)	−13.5(56)	−13.5(56)
	1000	0.2(100)	−3.6(79)	−4.0(79)	−4.0(79)	−4.0(79)	−4.0(79)

Table 5.4.2: Monte Carlo percent relative bias relative efficiency (in parentheses) of several estimators

5.4.2 Estimation of the mean square error

We also evaluated the performance of the proposed estimator of the mean square error given by (5.3.1) as well as the estimator of the mean square error for the once-winzorised estimator proposed by Rivest (1994, p. 380). We used a subset of the distributions considered in Section 5.4.1 consisting of the normal, exponential, Weibull and Pareto distributions. The number of samples and the sample sizes were identical to the ones used in Section 5.4.1. In this section, the estimator given $\overline{X}^R(c_{opt})$ was was based on the conditional bias estimated using the sample mean.

As a measure of the bias of the estimator of the mean square error, we used the Monte Carlo percent relative bias given by

$$RB_{MC}\left\{\widehat{MSE}\left(\widehat{\mu}\right)\right\} = \frac{E_{MC}\left\{\widehat{MSE}\left(\widehat{\mu}\right)\right\} - MSE_{MC}(\widehat{\mu})}{MSE_{MC}(\widehat{\mu})} \times 100,$$

where

$$MSE_{MC}(\widehat{\mu}) = \frac{1}{J}\sum_{j=1}^{J}\left(\widehat{\mu}^{(j)} - \mu\right)^2.$$

The results, presented in Table 5.4.3, show that the proposed estimator of the mean square error of $\overline{X}^R\left(c_{opt}\right)$ performed very well in terms of relative bias in all the scenarios with a an absolute relative bias less than 1%. The same was true for the estimator of the mean square error of $\overline{X}_1^R$ proposed by Rivest (1994).

	Normal		Exponential		Weibull		Pareto	
Sample size	$\overline{X}_1^R$	$\overline{X}^R(c_{opt})$	$\overline{X}_1^R$	$\overline{X}^R(c_{opt})$	$\overline{X}_1^R$	$\overline{X}^R(c_{opt})$	$\overline{X}_1^R$	$\overline{X}^R(c_{opt})$
25	0.1	−0.1	−0.1	0.1	−0.8	−0.4	−0.8	−0.2
50	−0.2	−0.2	−0.2	−0.1	−0.9	−0.6	−0.2	0.1
100	0.2	0.2	0.1	0.2	−0.6	−0.7	0.0	0.2
1000	0.5	0.5	0.2	0.2	−0.8	−0.7	0.2	0.3

Table 5.4.3: Monte Carlo percent relative bias of estimators of the mean square error for the once-winsorized and the proposed robust estimator

Appendix

Proof of Proposition 5.1 :

$$\begin{aligned} B_i\left(\overline{X}\right) &= \mathbb{E}\left(\overline{X}|X_i\right) - \mu \\ &= \frac{1}{n}X_i + \mathbb{E}\left(\frac{1}{n}\sum_{j=1,j\neq i}^{n} X_j \,|\, X_i\right) - \mu \end{aligned}$$

Since the random variables Y_i are assumed to be independent, we have

$$\begin{aligned} B_i\left(\overline{X}\right) &= \frac{1}{n}X_i + \mathbb{E}\left(\frac{1}{n}\sum_{j=1,j\neq i}^{n} X_j\right) - \mu \\ &= \frac{1}{n}X_i + \frac{n-1}{n}\mu - \mu \\ &= \frac{1}{n}\left(X_i - \mu\right). \end{aligned}$$

□

Proof of Proposition 5.2: We have

$$\begin{aligned} \mathbb{E}\left\{\widehat{B}_i\left(\overline{X}\right) \,|\, X_i = x_i\right\} =& \mathbb{E}_F\left\{\frac{1}{n}\left(X_i - \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} X_j\right) \,|\, X_i = x_i\right\} \\ =& \frac{1}{n}x_i - \frac{1}{n}\frac{1}{(n-1)}\left(n-1\right)\mu \\ =& \frac{1}{n}\left(x_i - \mu\right). \end{aligned}$$

□

Proof of Proposition 5.4 : Using an inequality proposed by David et al. (1970, p. 47) on the expectation of the minimum and the maximum from n independent and identically distributed random variables drawn from a continuous

distribution F with finite mean μ and finite variance σ^2, we have

$$\mu - \frac{(n-1)\sigma}{\sqrt{2n-1}} \leq \mathbb{E}\left(X_{(1)}\right) \leq \mathbb{E}\left(X_{(n)}\right) \leq \mu + \frac{(n-1)\sigma}{\sqrt{2n-1}}.$$

This leads to

$$\mu - \frac{(n-1)\sigma}{\sqrt{2n-1}} \leq \frac{1}{2}\left(\mu_1 + \mu_n\right) \leq \mu + \frac{(n-1)\sigma}{\sqrt{2n-1}}.$$

Thus,

$$-\frac{\sigma}{\sqrt{2n-1}} \leq Bias\left\{\overline{X}^R\left(c_{opt}\right)\right\} \leq \frac{\sigma}{\sqrt{2n-1}}$$

□

Proof of Proposition 5.5 : The error of the robust estimator $\overline{X}^R\left(c_{opt}\right)$ is given by

$$\overline{X}^R\left(c_{opt}\right) - \mu = \frac{n}{n-1}\overline{X} - \mu - \frac{1}{2(n-1)}\left(X_{(1)} + X_{(n)}\right).$$

Using the triangle inequality twice, we have

$$\begin{aligned}\left|\overline{X}^R(c_{opt}) - \mu\right| &\leq \left|\frac{n}{n-1}\overline{X} - \mu\right| + \frac{1}{2(n-1)}\left(\left|X_{(1)}\right| + \left|X_{(n)}\right|\right) \\ &\leq \frac{n}{n-1}\left|\overline{X} - \mu\right| + \frac{1}{n-1}\left|\mu\right| + \frac{1}{2(n-1)}\left(\left|X_{(1)}\right| + \left|X_{(n)}\right|\right).\end{aligned}$$

Taking expectation on both sides, we obtain

$$\begin{aligned}\mathbb{E}\left\{\left|\overline{X}^R(c_{opt}) - \mu\right|\right\} &\leq \frac{n}{n-1}\mathbb{E}\left\{\left|\overline{X} - \mu\right|\right\} + \frac{1}{n-1}|\mu| + \frac{1}{2(n-1)}\mathbb{E}\left\{\left|X_{(1)}\right| + \left|X_{(n)}\right|\right\} \\ &\leq \frac{n}{n-1}\mathbb{E}\left\{\left|\overline{X} - \mu\right|\right\} + \frac{1}{n-1}|\mu| + \frac{1}{(n-1)}\mathbb{E}\left\{\max\left(\left|X_{(1)}\right|, \left|\left|X_{(n)}\right|\right)\right\}.\end{aligned}$$

Using the inequality of David et al. (1970, p. 47) applied to the sample of the absolute value,$(|X_1|, ..., |X_n|)$ drawn from a continuous distribution $\tilde{F}$ with finite

mean $\tilde{\mu}$ and finite variance $\tilde{\sigma}^2$, we have

$$0 \leq \mathbb{E}_F\left[\max\left(|X_{(1)}|,..,|X_{(n)}|\right)\right] \leq \tilde{\mu} + \frac{(n-1)\tilde{\sigma}}{\sqrt{2n-1}}.$$

Therefore,

$$\mathbb{E}\left\{\left|\overline{X}^R(c_{opt}) - \mu\right|\right\} \leq \frac{n}{n-1}\mathbb{E}\left\{\left|\overline{X} - \mu\right|\right\} + \frac{1}{n-1}|\mu| + \frac{\tilde{\mu}}{n-1} + \frac{\tilde{\sigma}}{\sqrt{2n-1}}$$

Finally, letting $n \to +\infty$, we obtain the result.

□

Proof of Theorem 5.1 : First, we have

$$\begin{aligned} MSE\left\{\overline{X}^R(c_{opt})\right\} &= \mathbb{E}\left\{\left(\overline{X}^R(c_{opt}) - \mu\right)^2\right\} \\ &= \mathbb{E}\left\{\left(\frac{n}{n-1}\overline{X} - \mu\right)^2\right\} - \frac{n}{(n-1)^2}\mathbb{E}\left\{\left(\overline{X} - \mu\right)X_{(1)}\right\} \qquad (A.1) \\ &- \frac{n}{(n-1)^2}\mathbb{E}\left\{\left(\overline{X} - \mu\right)X_{(n)}\right\} \\ &- \frac{1}{(n-1)^2}\mathbb{E}\left\{\mu\left(X_{(1)} + X_{(n)}\right)\right\} + \frac{1}{4\left(n-1\right)^2}\mathbb{E}\left\{\left(X_{(1)} + X_{(n)}\right)^2\right\}. \end{aligned}$$

Now, the three terms $\mathbb{E}\left\{\left(\frac{n}{n-1}\overline{X} - \mu\right)^2\right\}$, $\mathbb{E}\left\{\mu\left(X_{(1)} + X_{(n)}\right)\right\}$ and $\mathbb{E}\left\{\left(X_{(1)} + X_{(n)}\right)^2\right\}$ are easy to derive:

$$\mathbb{E}\left\{\left(\frac{n}{n-1}\overline{X} - \mu\right)^2\right\} = \frac{1}{(n-1)^2}\left(\mu^2 + n\sigma^2\right), \qquad (A.2)$$

$$\mathbb{E}\left\{\mu\left(X_{(1)} + X_{(n)}\right)\right\} = \mu\left(\mu_1 + \mu_n\right) \qquad (A.3)$$

and

$$\mathbb{E}\left\{\left(X_{(1)} + X_{(n)}\right)^2\right\} = \left(\mu_{1,1} + \mu_{n,n} + 2\mu_{1,n}\right). \qquad (A.4)$$

We now focus on the covariance term $\mathbb{E}\left\{\left(\overline{X}-\mu\right)X_{(n)}\right\}$, which can be rewritten as

$$\mathbb{E}\left\{\left(\overline{X}-\mu\right)X_{(n)}\right\}=\sum_{i=1}^{n}\mathbb{E}\left(X_{(i)}X_{(n)}\right)-\mu n\mu_n.$$

Using a relationship among moments of order statistics given by Rivest (1994), we have

$$\forall i\in[1:n-1],\ \sum_{j=1}^{n}\sum_{k=i+1}^{n}\mu_{jk}=\sum_{j=i+1}^{n}(\mu_{jj}-\mu_{ij})+\mu(n-i)\mu_i+\mu(n-1)\sum_{j=i+1}^{n}\mu_j.$$

Taking $i=n-1$ in the previous expression, we obtain

$$\sum_{j=1}^{n}\mathbb{E}\left(X_{(j)}X_{(n)}\right)=\sum_{j=1}^{n}\mu_{j,n}=\mu_{n,n}-\mu_{n-1,n}+\mu\left(\mu_{n-1}+(n-1)\mu_n\right)$$

Thus,

$$\mathbb{E}\left\{\left(\overline{X}-\mu\right)X_{(n)}\right\}=\left[\mu_{n,n}-\mu_{n-1,n}+\mu\left(\mu_{n-1}-\mu_n\right)\right] \tag{A.5}$$

Likewise, by considering the original sample with an opposite sign, $(-X_1,...,-X_n)$, it can be shown that

$$\sum_{j=1}^{n}\mathbb{E}\left(X_{(j)}X_{(1)}\right)=\sum_{j=1}^{n}\mu_{j,1}=\mu_{1,1}-\mu_{2,1}+\mu\left(\mu_2+(n-1)\mu_1\right).$$

Then,

$$\mathbb{E}\left\{\left(\overline{X}-\mu\right)X_{(1)}\right\}=\mu_{1,1}-\mu_{2,1}+\mu\left(\mu_2-\mu_1\right). \tag{A.6}$$

Replacing the terms (A.2), (A.3), (A.4), (A.5) and (A.6) in (A.1), the mean square

error reduces to

$$\begin{aligned} MSE\left\{\overline{X}^R(c_{opt})\right\} &= \frac{1}{(n-1)^2}\left(\mu^2+n\sigma^2\right)-\frac{1}{(n-1)^2}\mu\left(\mu_1+\mu_n\right) \\ &- \frac{1}{(n-1)^2}\left[\mu_{n,n}-\mu_{n-1,n}+\mu\left(\mu_{n-1}-\mu_n\right)\right] \\ &- \frac{1}{(n-1)^2}\left[\mu_{1,1}-\mu_{2,1}+\mu\left(\mu_2-\mu_1\right)\right] \\ &+ \frac{1}{4\left(n-1\right)^2}\left(\mu_{1,1}+\mu_{n,n}+2\mu_{1,n}\right). \end{aligned}$$

This leads to the expression given in Theorem 5.1.

□

Proof of Corollary 5.1 : We have

$$\begin{aligned} MSE\left\{\overline{X}^R(c_{opt})\right\} &= \mathbb{E}\left\{\left(\overline{X}^R-\mu\right)^2\right\} \\ &= \mathbb{E}\left\{\left(\frac{n}{n-1}\overline{X}-\mu\right)^2\right\}+\frac{1}{(n-1)^2}\mathbb{E}\left\{\left(\frac{X_{(1)}+X_{(n)}}{2}\right)^2\right\} \\ &- \frac{1}{n-1}\mathbb{E}\left\{\left(\frac{n}{n-1}\overline{X}-\mu\right)\left(X_{(1)}+X_{(n)}\right)\right\} \\ &= \mathbb{E}\left\{\left(\frac{n}{n-1}\overline{X}-\mu\right)^2\right\}+\frac{1}{(n-1)^2}\left\{\mathbb{V}\left(\frac{X_{(1)}+X_{(n)}}{2}\right)+\mathbb{E}\left(\frac{X_{(1)}+X_{(n)}}{2}\right)^2\right\} \\ &- \frac{n}{(n-1)^2}\mathbb{E}\left\{\left(\overline{X}-\mu\right)\left(X_{(1)}+X_{(n)}\right)\right\}-\frac{1}{(n-1)^2}\mu\mathbb{E}\left\{\left(X_{(1)}+X_{(n)}\right)\right\} \end{aligned} \tag{A.7}$$

Since the normal distribution is symmetric, we have

$$\mathbb{E}\left\{\frac{\left(X_{(1)}+X_{(n)}\right)}{2}\right\}=\mu. \tag{A.8}$$

Furthermore, for a normal distribution with finite mean μ, the independence of

$X_{(r)} - \overline{X}$ and $\overline{X} - \mu$ for any $r \in [1:n]$ leads to

$$\mathbb{E}\left[\left\{\left(\overline{X} - \mu\right)\right\} X_{(r)}\right] = 0.$$

Therefore,

$$\mathbb{E}\left[\left\{\left(\overline{X} - \mu\right)\right\} \left(X_{(1)} + X_{(n)}\right)\right] = 0. \tag{A.9}$$

Replacing the terms (A.2), (A.8) and (A.9) in (A.7), the mean square error reduces to

$$\begin{aligned} MSE\left\{\overline{X}^R(c_{opt})\right\} &= \frac{1}{(n-1)^2}\left(\mu^2 + n\sigma^2\right) + \frac{1}{(n-1)^2}\left\{\mathbb{V}\left(\frac{X_{(1)} + X_{(n)}}{2}\right) + \mu^2\right\} - 2\frac{\mu^2}{(n-1)^2} \\ &= \frac{1}{(n-1)^2}\left\{(n-1)\,\sigma^2 + \mathbb{V}\left(\frac{X_{(1)} + X_{(n)}}{2}\right)\right\}. \end{aligned}$$

Using an approximation of the variance of the mid-range given by Cramer (1946,p 376), we have

$$\mathbb{V}(M) = \frac{\sigma^2\pi^2}{24log(n)} + O\left(\frac{1}{log(n)^2}\right).$$

Finally,

$$MSE\left\{\overline{X}^R(c_{opt})\right\} = \frac{n}{(n-1)}\left[\frac{\sigma^2}{n} + \frac{\sigma^2\pi^2}{24n(n-1)log(n)} + O\left\{\frac{1}{n^2log(n)^2}\right\}\right].$$

□

Proof of Corollary 5.2 : Approximation of $MSE\left\{\overline{X}^R(c_{opt})\right\}$ for distributions belonging to the max-domain of attraction Gumbel: $\Lambda_3(x) = \exp(-\exp(-x))$.

To approximate MSE of $\overline{X}^R(c_{opt})$, we regard the ordered sample from F, $\left(X_{(1)}, .., X_{(n)}\right)$ as a function of the ordered sample from the exponential distribution $\left\{h\left(Y_{(1)}\right), .., h\left(Y_{(n)}\right)\right\}$, where $h(t) = F^{-1}\left(1 - e^{-t}\right)$. Then, the vector $\left(X_{(n-1)}, X_{(n)}\right)$ has the same distribution as $\left\{h\left(Y_{(n-1)}\right), h\left(Y_{(n)}\right)\right\}$. Using the gen-

eralized Rényi's representation of an exponential order statistics as a linear function of independent exponentials, we have

$$Y_{(n)} - Y_{(n-1)} = V_n,$$

with $V_n \sim \mathcal{E}(1)$.

It follows that $\left(X_{(n-1)}, X_{(n)}\right)$ has the same distribution as $\left\{h\left(Y_{(n-1)}\right), h\left(Y_{(n-1)} + V_n\right)\right\}$.

Now, we lay out a Taylor expansion of the function h.

If F belongs to the max-domain of attraction Gumbel and satisfies the von Mises condition (David, 1970, p. 207), we have $\lim_{x\to+\infty} \frac{d}{dx}\left[\frac{1-F(x)}{f(x)}\right] = 0$, i.e $\lim_{x\to+\infty} \frac{f^{'}(x)(1-F(x))}{f(x)^2} = -1$ using L'Hôpital's rule, we have : $\lim_{x\to+\infty} \frac{1-F(x)}{xf(x)} = 0$. This leads to the third-order Taylor expansion of $h(t+x)$=$F^{-1}(1-e^{-t-x})$,

$$\frac{h(t+x)}{h(t)} = 1 + xg_1(t) + x^2 g_2(t) + O(x^3 g_1^3(t)), \tag{A.10}$$

where

$$g_1(t) = \left[e^t f\left\{F^{-1}\left(1-e^{-t}\right)\right\} F^{-1}(1-e^{-t})\right]^{-1} = \frac{e^{-t}}{f\left\{F^{-1}\left(1-e^{-t}\right)\right\} F^{-1}(1-e^{-t})},$$

$$2g_2(t) = -g_1(t) - g_1^2(t)\frac{f^{'}\left\{F^{-1}(1-e^{-t})\right\} F^{-1}\left(1-e^{-t}\right)}{f\left\{F^{-1}\left(1-e^{-t}\right)\right\}},$$

and $g_1(t)$ and $g_2(t)$ tend to 0 as t goes to infinity.

Assuming that F is strictly increasing and using the substitution $z = F^{-1}\left(1-e^{-t}\right)$, we have

$$g_1(z) = \frac{1-F(z)}{zf(z)}.$$

Thus,

$$\lim_{t\to+\infty} g_1(t) = \lim_{z\to+\infty} \frac{1-F(z)}{zf(z)} = 0.$$

Likewise,

$$2g_2(t) = -g_1(t) - g_1^2(t)\frac{f'\left\{F^{-1}(1-e^{-t})\right\}F^{-1}\left(1-e^{-t}\right)}{f\left\{F^{-1}\left(1-e^{-t}\right)\right\}},$$

using the substitution $z = F^{-1}\left(1-e^{-t}\right)$, we have

$$\begin{aligned}
2g_2(z) &= -g_1(z) - g_1^2(z)\frac{zf'(z)}{f(z)} \\
&= -g_1(z) - \left(\frac{1-F(z)}{zf(z)}\right)^2 \frac{zf'(z)}{f\{z\}} \\
&= -g_1(z) - \frac{1-F(z)}{zf(z)}\frac{\left[1-F(z)\right]f'(z)}{f^2(z)}.
\end{aligned}$$

Using the von Mises condition (David, 1970, p. 207), we have

$$\lim_{z\to+\infty}\frac{1-F(z)}{zf(z)} = 0$$

and

$$\lim_{z\to+\infty}\frac{\left[1-F(z)\right]f'(z)}{f^2(z)} = -1,$$

we obtain

$$\lim_{t\to+\infty} g_2(t) = \lim_{z\to+\infty} g_2(z) = 0.$$

Using the Taylor expansion (A.10), the vector of the two largest observations $\left(X_{(n-1)}, X_{(n)}\right)$ has the same distribution as

$$h\left(Y_{(n-1)}\right)\left[1 + g_1\left(Y_{(n-1)}\right)\begin{pmatrix}0\\V_n\end{pmatrix} + g_2\left(Y_{(n-1)}\right)\begin{pmatrix}0\\V_n^2\end{pmatrix} + O\left\{g_1^3\left(Y_{(n-1)}\right)\right\}\right]$$

For example, $X_{(n)}\left(X_{(n)}-X_{(n-1)}\right)$ has the same distribution as

$$h^2\left\{log\left(Y_{(n-1)}\right)\right\}\left[g_1\left\{log\left(Y_{(n-1)}\right)\right\}V_n\right]+O\left[h^2\left\{log\left(Y_{(n-1)}\right)\right\}g_2\left\{log\left(Y_{(n-1)}\right)\right\}\right].$$

It follows that

$$\begin{aligned}\mathbb{E}\left\{X_{(n)}\left(X_{(n)}-X_{(n-1)}\right)\right\} &= \mathbb{E}\left[\mathbb{E}\left\{X_{(n)}\left(X_{(n)}-X_{(n-1)}\right)|Y_{(n-1)}\right\}\right]\\ &= \mathbb{E}\left\{\mathbb{E}\left(h^2\left(Y_{(n-1)}\right)\left[g_1\left(Y_{(n-1)}\right)V_n+g_1^2\left(Y_{(n-1)}\right)V_n^2+g_2\left(Y_{(n-1)}\right)V_n^2\right.\right.\\ &+ \left.\left.2g_1\left(Y_{(n-1)}\right)g_2\left(Y_{(n-1)}\right)V_n^3+O\left\{g_1^3\left(Y_{(n-1)}\right)\right\}|Y_{(n-1)}\right]\right)\right\}.\end{aligned}$$

Since in large sample, we have $Y_{(n-1)}=log(n)+O_p(1)$, it follows that

$$\begin{aligned}\mathbb{E}\left\{X_{(n)}\left(X_{(n)}-X_{(n-1)}\right)\right\} &= h^2\left\{log(n)\right\}\left(g_1\left\{log(n)\right\}+2\left[g_1^2\left\{log(n)\right\}+g_2\left\{log(n)\right\}\right]\right.\\ &+ \left.12g_1\left\{log(n)\right\}g_2\left\{log(n)\right\}\right)+O\left[h^2\left\{log(n)\right\}g_1^3\left\{log(n)\right)\right\}\right]\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left(X_{(n-1)}\right) &= \mathbb{E}\left\{\mathbb{E}\left(X_{(n-1)}|Y_{(n-1)}\right)\right\}\\ &= h\left\{log(n)\right\}+O(1).\end{aligned}$$

Finally,

$$\mu_{n,n}-\mu_{n-1,n}=h^2\left\{log(n)\right\}\left[g_1\left\{log(n)\right\}+2g_1^2\left\{log(n)\right\}\right]+O\left[h^2\left\{log(n)\right\}g_2\left\{log(n)\right\}\right]. \tag{A.11}$$

Likewise, we can show that

$$\begin{aligned}\mu_{n,n} &= h^2\{log(n)\}\left[1+2g_1\{log(n)\}+2g_1^2\{log(n)\}+\right.\\ &+ \left. 12g_1\{log(n)\}g_2\{log(n)\}+4g_2\{log(n)\}\right]\\ &+ O\left[h^2\{log(n)\}g_1^3\{log(n)\}\right] \qquad (A.12)\end{aligned}$$

and

$$\mu_n = h\{log(n)\}\left[1+g_1\{log(n)\}+2g_2\{log(n)\}\right]+O\left[h\{log(n)\}g_1^3\{log(n)\}\right]. \qquad (A.13)$$

Assuming that the support of the distribution is bounded below by w_{min}, we have

$$\mu_1 = O(1); \mu_2 = O(1); \mu_{11} = O(1). \qquad (A.14)$$

It remains to study the term

$$\mu_{1n} = \mathbb{E}_F\left(X_{(1)}X_{(n)}\right).$$

Since $X_{(1)}$ and $X_{(n)}$ are asymptotically independent in the sense that

$$\left|F_{X_{(1)},X_{(n)}}(x,y) - F_{X_{(1)}}(x)F_{,X_{(n)}}(y)\right| < \delta,\ \forall n > 15$$

provided $\delta \in [0.00933; 0.01001]$ (see Walsh (1970) for more details), we have

$$\begin{aligned}\mu_{1n} &\backsimeq \mathbb{E}\left(X_{(1)}\right)\mathbb{E}\left(X_{(n)}\right)\\ &\backsimeq -\left\{w_{min}-\mathbb{E}\left(X_{(1)}\right)\right\}\mathbb{E}\left(X_{(n)}\right)+w_{min}\mathbb{E}\left(X_{(n)}\right).\end{aligned}$$

Assuming that the right tail distribution on $[w_{min}; a]$,$a \in \mathbb{R}$ has the form $F(x) = K\left(w_{min} - x\right)^{\alpha}$ with $\alpha > 0$, we have

$$(Kn)^{1/\alpha}\left(X_{(1)} - w_{min}\right) \longrightarrow \overline{\Psi}_{\alpha},$$

where

$$\overline{\Psi}_{\alpha}(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - exp\left(-x^{\alpha}\right) & \text{si } x > 0 \end{cases}$$

Since the random variable $(X_1, ...X_n)$ have a finite second moment, we can use the Theorem on moment convergence of maximum (see de Haan, p. 177) to the moment of the limit distribution of the maximum

$$\lim_{n \to +\infty} \mathbb{E}\left\{(Kn)^{1/\alpha}\left(X_{(1)} - w_{min}\right)\right\} = \mathbb{E}\left(Z\right),$$

where $Z \sim \overline{\Psi}_{\alpha}$. It follows that

$$\mathbb{E}\left(X_{(1)} - w_{min}\right) = O\left(\frac{1}{n^{\beta}}\right),$$

where $0 < \beta < \frac{1}{\alpha}$. Hence, we have

$$\begin{aligned} \mu_{1n} &= \left\{\mathbb{E}\left(X_{(1)}\right) - w_{min}\right\}\mathbb{E}\left(X_{(n)}\right) + w_{min}\mathbb{E}\left(X_{(n)}\right) \\ &= w_{min}\mathbb{E}_F\left(X_{(n)}\right) + o(1). \end{aligned}$$

Using the expression (A.13) for the expectation of the maximum, we obtain

$$\mu_{1n} = w_{min} h\left\{log(n)\right\}\left[1 + g_1\left\{log(n)\right\} + 2g_2\left\{log(n)\right\}\right] + O\left[h\left\{log(n)\right\} g_1^3\left\{log(n)\right\}\right]. \tag{A.15}$$

Now, we show that the term $O\left[h\left\{log(n)\right\} g_1^3\left\{log(n)\right\}\right]$ is negligible compared to

the term $O\left[h^2\left\{log(n)\right\}g_1^3\left\{log(n)\right\}\right]$.

$$\frac{h(t)g_1^3(t)}{h^2(t)g_1^3(t)} = \frac{1}{h(t)}.$$

Since

$$\lim_{t\to\infty}\frac{1}{h(t)} = 0,$$

we have

$$\lim_{t\to\infty}\frac{h(t)g_1^3(t)}{h^2(t)g_1^3(t)} = 0.$$

Replacing the terms (A.11), (A.12) and (A.15), in the general expression of the mean square error given in Theorem 7, we obtain the following $O\left[\frac{h^2\{log(n)\}g_1^3\{log(n)\}}{n^2}\right]$ approximation :

$$\begin{aligned}
MSE\left\{\overline{X}^R(c_{opt})\right\} =& \frac{1}{(n-1)^2}\left\{n\sigma^2 - h^2\left\{log(n)\right\}\left(g_1\left\{log(n)\right\} + 2\left[g_1^2\left\{log(n)\right\} + g_2\left\{log(n)\right\}\right]\right.\right.\\
&+ 12g_1\left\{log(n)\right\}g_2\left\{log(n)\right\}) - \mu h\left\{log(n)\right\}\}\\
&+\frac{1}{4(n-1)^2}\left(h^2\left\{log(n)\right\}\left[1 + 2g_1\left\{log(n)\right\} + 2g_1^2\left\{log(n)\right\}\right.\right.\\
&+ 12g_1\left\{log(n)\right\}g_2\left\{log(n)\right\} + 4g_2\left\{log(n)\right\}]\\
&+ 2w_{min}h\left\{log(n)\right\}\left[1 + g_1\left\{log(n)\right\} + 2g_2\left\{log(n)\right\}\right])\\
&+O\left[\frac{h^2\left\{log(n)\right\}g_1^3\left\{log(n)\right\}}{n^2}\right], \qquad \text{(A.16)}
\end{aligned}$$

Proof of Corollary 5.3 : Approximation of $MSE\left\{\overline{X}^R(c_{opt})\right\}$ for the Pareto distribution with $F(x) = 1 - x^{-\gamma}$. Since the support of the Pareto distribution is bounded below, we have

$$\mu_1 = O(1); \mu_2 = O(1); \mu_{11} = O(1).$$

The expression of the mean square error in Theorem 5.1 reduces to

$$\begin{aligned} MSE\left\{\overline{X}^{R}\left(c_{opt}\right)\right\} &= \frac{n\sigma^2}{(n-1)^2} - \frac{1}{(n-1)^2}\left[\mu_{n,n} - \mu_{n-1,n} + \mu\mu_{n-1}\right] + \frac{1}{4\left(n-1\right)^2}\left(\mu_{n,n} + 2\mu_{1,n}\right) \\ &+ O\left(\frac{1}{n^2}\right). \end{aligned}$$

By replacing the moments of order statistics by their expression using the Gamma function given by Huang (1975)

$$\mu_n = n!\frac{\Gamma(1-1/\gamma)}{\Gamma(n+1-1/\gamma)},$$

$$\mu_{n,n} = n!\frac{\Gamma\left(1-2/\gamma\right)}{\Gamma\left(n+1-2/\gamma\right)},$$

$$\mu_{1,n} = n!\frac{\Gamma\left(n-2/\gamma\right)\Gamma\left(1-1/\gamma\right)}{\Gamma\left(n+1-2/\gamma\right)\Gamma\left(n-1/\gamma\right)}$$

and using the following relationship among moments of order statistics for the Pareto distribution

$$\mu_{n-1} = \left(1-\frac{1}{\gamma}\right)\mu_n,$$

$$\mu_{n,n-1} = \frac{\gamma-2}{\gamma-1}\mu_{n,n},$$

we obtain

$$\mu_{n,n} - \mu_{n,n-1} = \frac{1}{\gamma-1}\mu_{n,n}.$$

Tedious but straightforward calculations yield the following $O\left[\frac{1}{n^2}\right]$ approximation :

$$MSE\left\{\overline{X}^{R}\left(c_{opt}\right)\right\} = \frac{n\sigma^2}{(n-1)^2} + \frac{1}{(n-1)^2}\left(\frac{1}{4} - \frac{1}{\gamma-1}\right)\mu_{n,n} + \frac{1}{(n-1)^2}\left\{\frac{\mu_{1,n}}{2} - \mu\left(1-\frac{1}{\gamma}\right)\mu_n\right\}.$$

□

References

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555–569.

Cramer,H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

David, H. A. and Nagaraja, H. N. (1970). *Order statistics.* Wiley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383–393.

Huang, J. (1975). A note on order statistics from Pareto distribution. *Scandinavian Actuarial Journal*, 1975(3), 187–190.

Maronna, R. A. R. D., Martin, D., & Yohai, V. (2006). *Robust statistics* .Wiley.

Muñoz-Pichardo, J., Muñoz-Garcia, J., Moreno-Rebollo, J.L. and Piño-Mejias, R. (1995). A new approach to influence analysis in linear models. *Sankhya Series A,* 57, 393–409.

Myers, R. and Pepin, P. (1990). The robustness of lognormal-based estimators of abundance. *Biometrics*, 46, 1185–1192.

Rivest, L.-P. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika*, 81, 373–383.

Sarhan, A. E. and Greenberg, B. G. (1962). *Contributions to order statistics.* Wiley.

Searls, D. T. (1966). An estimator for a population mean which reduces the effect of large true observations. *Journal of the American Statistical Association*, 61, 1200–1204.

Walsh, J.E. (1970). Sample sizes for approximate independence of largest and smallest order statistics. *Journal of the American Statistical Association, 65, 860–863.*

Chapter 6

Robust prediction for GLM and GLMM in finite population

Résumé

En pratique, il est courant d'observer des unités influentes dans les échantillons collectés, plus particulièrement lorsque l'on collecte de l'information sur des variables économiques dont la distribution est très asymétrique. Le fait d'ajouter ou de retirer cette unité dite influente a un impact significatif sur les predicteurs classiquement utilisés pour inférer sur des paramètres de population finie. La présence d'unités influentes est d'autant plus néfaste que la taille de l'échantillon est petite, c'est pourquoi les méthodes de prediction robuste sur petits domaines se sont développées de façon importante au cours de ses dernières années, voir par exemple Gosh et al. (2008), Sinha and Rao (2009), Dongmo Jiongo et al. (2013), Chambers et al. (2013) et Fabrizi et al. (2014). La majorité de ces travaux reposent sur l'utilisation de modèle mixte au niveau des unités et s'intéressent à des variables d'intérêt continues. Dans ce cadre, quelques versions robustes du predicteur linéaire sans biais optimal empirique ont été proposés dans la littérature en utilisant des méthodes de type M-quantile ou une approche basée sur le biais conditionnel. En pratique, il est courant de s'intéresser à des variables d'intérêt binaires ou discrètes. On a alors recours à des modèles logistiques mixtes ou des modèles de Poisson mixtes. Nous proposons dans un premier temps un predicteur robuste dans le cas d'une approche sous le modèle avec utilisation d'un modèle GLM, puis nous proposons une approche unifiée pour la prediction robuste dans les petits domaines dans le cadre des modèles GLMM.

Mots clés : Biais conditionnel ; prediction robuste ; approche modèle ; petits domaines.

Abstract

Influential units occur frequently in surveys, especially in the context of business surveys that collect economic variables whose distribution are highly skewed. A unit is said to be influential when its inclusion or exclusion from the sample has an important impact on the magnitude of survey statistics. Robust small area prediction has received a lot of attention in recent years; see Gosh et al. (2008), Sinha and Rao (2009), Dongmo Jiongo et al. (2013), Chambers et al. (2013) and Fabrizi et al. (2014), among others. So far, researchers have mainly focused on unit level models and continuous characteristics of interest. Several robust versions of the empirical best linear unbiased predictor based on linear mixed models (LMM) have been proposed in the literature, including an M-quantile regression approach and an approach based on the concept of conditional bias of a unit. In practice, one must often face binary and count data. In this case, methods based on LMMs are not suited. We first propose a robust predictor in a general model-based framework with the use of generalized linear models and then we propose a unified framework for robust small area prediction in the context of generalized LMMs. We construct a general robust predictor based on the concept of conditional bias.

Keywords: Conditional bias; robust prediction; model-based approach; small-areas .

Introduction

Influential units are common feature of many sample surveys, especially in the context of business surveys, where the variables of interest have generally highly skewed distribution. There exists two main inferential approaches in finite populations sampling : the design-based also called "the randomisation approach" and the model-based approach also called the prediction approach. In this chapter, we focus on the latter approach. In this context, assuming linear relationship between the dependent variable and a vector of predictors, a frequently used predictor of a finite population total is the Empirical Best Linear Unbiased Predictor (EBLUP); for example see Dorfmann et al. (2000) and Chambers et al. (2012) . In the presence of influential units, the EBLUP remains unbiased but its variance can be very large and some robust estimators have already been proposed, see Chambers (1986), Clark (1995) and Beaumont et al. (2013) among others. The model-based approach is widely used for small estimation to overcome the sample size problem. A small area is defined as a domain whose sample size is too small to obtain a reliable direct estimator. In this context, a linear mixed model is assumed, leading to the Empirical Best Unbiased Predictor (EBP). Some robust estimators have been proposed for unit-level model based on a linear mixed model; see Gosh et al. (2008), Sinha and Rao (2009), Dongmo Jiongo et al. (2013), Chambers et al. (2014) and Fabrizi et al. (2014), among others. So far, researchers have mainly focused on unit-level models and continuous characteristics of interest. In practice, one must often face binary and count data, which requires the use of generalized linear mixed models. For example, for a binary outcome in a frequentist approach, one can follow Jiang and Lahiri (2001), or Jiang (2003) who propose an Empirical Best Predictor for generalized linear mixed models (GLMM). These estimators are also very sensitive to the presence of influential units, which is why it is desirable to develop robust prediction in this context. To our knowledge, this problem has not been adressed in the literature with two notable exeptions,

Tzavidis et al. (2014) for count data and Chambers et al.(2014) for binary data using M-quantile regression approach. In this chapter we start by extending the results of Beaumont et al. (2013) to the case of GLM. We assess empirically the properties of the proposed robust estimator in terms of bias and efficiency. Then, we consider the problem of generalized linear mixed model in the special case of small area estimation with a empirical comparison to the M-quantile estimator.

6.1 Robust estimation for GLM

6.1.1 Model-based approach using GLM

Let U denote a finite population of size N. Let Y_i be the variable of interest attached to unit i. In model-based inference for finite population sampling (see e.g., Valliant, Dorfman and Royall, 2000), the y-values of the N population units are assumed to be generated from a model. The interest lies in the prediction of a function of the population $y-$values through the sample y-values. We assume that the Y_i's are independent random variable and that their distribution belongs to the exponential family. We denote by $\boldsymbol{X}$, the known N-row matrix containing the vector of explanatory variables $\boldsymbol{x}_i^\top$ in its i-th row. We assume that the vector $\boldsymbol{x}_i^\top$ is recorded without error for all population units. A non-informative sample s is selected from the finite population U and is treated as fixed when making inferences. We are interested in predicting the random population total $\theta = \sum_{i\in U} Y_i$ using the auxiliary information $\boldsymbol{X}$.

We assume a generalized linear model (GLM) for $\mu_i = E_m(Y_i)$ of the form

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{6.1.1}$$

where $g(.)$ is the link function, assumed to be known and invertible.

Model (6.1.1) includes three important special cases :

(i) the linear model obtained with $g(.)$ equal to the identity function;

(ii) the Bernoulli-Logistic GLM, where $g(.)$ is the logistic link function and the individual Y_i values are taken to be independent Bernoulli outcomes with

$$\mu_i = E_m\left(Y_i\right) = \frac{exp(\eta_i)}{1 + exp(\eta_i)},$$

and

$$Var_m\left(Y_i\right) = \mu_i\left(1 - \mu_i\right);$$

(iii) the Poisson-log GLM where $g(.)$ is the log link function and the individual Y_i values are taken to be independent Poisson random variable with

$$\mu_i = E_m\left(Y_i\right) = Var_m\left(Y_i\right) = exp(\eta_i).$$

The minimum mean squared error predictor of $\theta = \sum_{i \in U} Y_i$ is

$$\theta^* = \sum_{i \in s} Y_i + \sum_{i \in U \setminus s} E_m\left(Y_i | Y_j = y_j,\ j \in s\right). \tag{6.1.2}$$

If the model (6.1.1) holds, the minimum mean squared error predictor θ^* reduces to

$$\theta^* = \sum_{i \in s} Y_i + \sum_{i \in U \setminus s} h\left(\eta_i\right),$$

where $h(.)$ is the inverse of the link function $g(.)$. The predictor θ^* is impossible to compute in practice since it depends on the unknown superpopulation parameter β. A prediction of θ^* is obtained by replacing the unknown parameter $\boldsymbol{\beta}$ with a suitable estimate $\hat{\boldsymbol{\beta}}$. This leads to an Empirical Best Predictor

$$\hat{\theta}^{EBP} = \sum_{i \in s} Y_i + \sum_{i \in U \setminus s} h\left(\hat{\eta}_i\right) = \sum_{i \in s} Y_i + \sum_{i \in U \setminus s} h\left(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\right). \tag{6.1.3}$$

Since a non-informative sampling design is assumed, a suitable estimator $\hat{\boldsymbol{\beta}}$ can be obtained as the solution of the sample estimating equation

$$\sum_{i \in s} (Y_i - \mu_i) w_i g_i^{'}(\mu_i) \boldsymbol{x}_i = 0, \tag{6.1.4}$$

where $w_i = \left\{v(\mu_i) g_i^{'}(\mu_i)\right\}^{-1}$, $v(\mu_i) = Var(Y_i)$ and $g_i^{'}(\mu_i) = \partial g(\mu_i)/\partial \mu_i = \partial \eta_i / \partial \mu_i$.

The Empirical Best Predictor is generally biased but is consistent under some mild conditions. However, the EBP may be unstable in the presence of influentials units. Thus, it seems desirable to provide a robust version of the Empirical Best Predictor using the concept of conditional bias introduced in the context of model-based inference by Beaumont et al. (2013). Our goal is to provide a robust version of $\hat{\theta}^{EBP}$ which is not much affected when the GLM holds but which exhibits a smaller mean square error in presence of influentials units. It may be tempting to replace $\boldsymbol{\beta}$ in (6.1.3) by a robust estimator $\hat{\boldsymbol{\beta}}^R$ (e.g., an M-estimator proposed in the GLM case by Cantoni and Ronchetti (2001)), leading to

$$\hat{\theta}^{RCantoni} = \sum_{i \in S} Y_i + \sum_{i \in U \setminus S} h\left(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^R(c_R)\right). \tag{6.1.5}$$

In the linear case, replacing $\boldsymbol{\beta}$ in (6.1.3) by a robust M-estimator $\hat{\boldsymbol{\beta}}^R$, leads to Chambers (1986) robust version of the BLUP. The main difficulty lies in the choice of the tuning constant c_R appearing in the M-estimator $\hat{\boldsymbol{\beta}}^R$. Huber (1981) argued that $c_R = 1.345$ is a good choice, and showed that asymptotically, it is 95% as efficient as least squares if the true distribution is normal. This tuning constant allows the robust methods to perform very well in a classical statistical

context, but its common knowledge that in case of a finite population this choice of the tuning constant leads to robust predictors which are too biased and do not perform very well. This fact will be illustrated empirically in the simulation study.

6.1.2 The conditional bias in GLM

As a measure of influence of a unit i on the Empirical Best Predictor $\hat{\theta}^{EBP}$, we consider the conditional bias of the unit i defined by

$$B_i^{EBP} = E_m\left(\hat{\theta}^{EBP} - \theta | s; Y_i = y_i\right). \tag{6.1.6}$$

The conditional bias depends on whether or not unit i is selected in the sample. Let I_i be the sample selection indicator variable for unit i such that $I_i = 1$ if $i \in S$ and $I_i = 0$, otherwise. We note by $B_i^{EBP}(I_i = 1)$ the conditional bias attached to the sampled unit i and by $B_i^{EBP}(I_i = 0)$ the conditional bias attached to the non-sampled unit i. Since the Empirical Best Predictor is no longer linear in the y-values, we use a first-order Taylor expansion, yielding the following approximation of the conditional bias :

$$B_i^{EBP} = \begin{cases} \sum_{j\in U\setminus s} h_j'(\eta_j)\boldsymbol{x}_j^\top E_m\left[\left\{\sum_{k\in s}\boldsymbol{H}\left(y_k,\boldsymbol{\beta}\right)\right\}^{-1}\sum_{k\in S}\boldsymbol{t}(y_k,\boldsymbol{\beta})\,|\,s, Y_i = y_i\right] & \text{if } i\in s \\ \sum_{j\in U\setminus s} h_j'(\eta_j)\boldsymbol{x}_j^\top E_m\left[\left\{\sum_{j\in s}\boldsymbol{H}\left(y_j,\boldsymbol{\beta}\right)\right\}^{-1}\sum_{k\in s}\boldsymbol{t}\left(y_k,\boldsymbol{\beta}\right)\right]\left\{y_i - h\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)\right\} & \text{if } i\in U\setminus s \end{cases} \tag{6.1.7}$$

where $\boldsymbol{t}(y_i,\boldsymbol{\beta}) = (Y_i - \mu_i)\, w_i g_i'(\mu_i)\boldsymbol{x}_i$ and $\boldsymbol{H}(y_i,\boldsymbol{\beta}) = \frac{\partial \boldsymbol{t}(y_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$. The details are given in the Appendix.

When the Hessian matrix $\sum_{j\in S}\boldsymbol{H}(y_j,\boldsymbol{\beta})$ does not depend on the y-values, which is the case if $g(.)$ is the canonical link function, the general expression of

the conditional bias (6.1.7) reduces to

$$B_i^{EBP} = \begin{cases} \sum_{j\in U\setminus s} h_j^{'}(\eta_j)\boldsymbol{x}_j^{\top}\left\{\sum_{k\in s}\boldsymbol{H}\left(\boldsymbol{\beta}\right)\right\}^{-1}\boldsymbol{x}_i w_i g_i^{'}(\mu_i)\left\{y_i - h\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)\right\} & \text{if } i\in s \\ -\left\{y_i - h(\boldsymbol{x}_i^{\top}\boldsymbol{\beta})\right\} & \text{if } i\in U\setminus s \end{cases} \tag{6.1.8}$$

The expression (6.1.8) show that a unit has a large influence if its residual $e_i = y_i - h\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)$ is large (which is often call vertical outlier) and if its $\boldsymbol{x}_i$-values are large compared to the Hessian matrix $\sum_{k\in s}\boldsymbol{H}\left(\boldsymbol{\beta}\right)$ (which is often call as horizontal outlier).

In the sequel, we restrict our attention to the case where $g(.)$ is the canonical link function. Next, we consider three special cases of (6.1.8).

Example 6.1. Linear case

Replacing $h(.) = I(.)$, $h_j^{'}(\eta_j) = 1$, $w_i g_i^{'}(\mu_i) = 1$ and $\sum_{k\in s}\boldsymbol{H}\left(\boldsymbol{\beta}\right) = \sum_{k\in s}\boldsymbol{x}_k\boldsymbol{x}_k^{\top}$ in the expression (6.1.8), we obtain

$$B_i^{EBP} = \begin{cases} \sum_{j\in U\setminus s}\boldsymbol{x}_j^{\top}\left(\sum_{k\in s}\boldsymbol{x}_k\boldsymbol{x}_k^{\top}\right)^{-1}\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right) & \text{if } i\in s \\ -\left(y_i - \boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right) & \text{if } i\in U\setminus s \end{cases}$$

which is identical to the expression of Beaumont et al. (2013, p.4).

Example 6.2. Logistic case

In the logistic case, we have

$$h(.) = \frac{exp\left(.\right)}{1+exp\left(;\right)},\ h_j^{'}(\eta_j) = \frac{exp\left(\boldsymbol{x}_j^{\top}\boldsymbol{\beta}\right)}{\left\{1+exp\left(\boldsymbol{x}_j^{\top}\boldsymbol{\beta}\right)\right\}^2},\ w_i g_i^{'}(\mu_i) = 1,$$

$$\sum_{k\in s} \boldsymbol{H}\left(y_k, \boldsymbol{\beta}\right) = \sum_{k\in s} \frac{exp\left(\boldsymbol{x}_k^\top \boldsymbol{\beta}\right)}{\left\{1 + exp\left(\boldsymbol{x}_k^\top \boldsymbol{\beta}\right)\right\}^2} \boldsymbol{x}_k \boldsymbol{x}_k^\top$$

and

$$\phi_i = \frac{exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\left\{1 + exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right\}}$$

The conditional bias in (6.1.8) reduces to

$$B_i^{EBP} = \begin{cases} \sum_{j\in U\setminus s} \phi_j \left(1-\phi_j\right) \boldsymbol{x}_j^\top \left[\sum_{k\in s} \phi_k \left(1-\phi_k\right) \boldsymbol{x}_k \boldsymbol{x}_k^\top\right]^{-1} \boldsymbol{x}_i \left[y_i - \phi_i\right] & \text{if } i \in s \\ -\left[y_i - \phi_i\right] & \text{if } i \in U\setminus s \end{cases}$$

Example 6.3. Poisson case

Replacing $h(.) = h^{'}(.) = exp\,(.)$, $w_i g_i^{'}(\mu_i) = 1$ and

$$\sum_{k\in s} \boldsymbol{H}\left(\boldsymbol{\beta}\right) = \sum_{k\in s} exp\left(\boldsymbol{x}_k^\top \boldsymbol{\beta}\right) \boldsymbol{x}_k \boldsymbol{x}_k^\top$$

in the expression (6.1.8), we obtain

$$B_i^{EBP} = \begin{cases} \sum_{j\in U\setminus s} exp\left(\boldsymbol{x}_j^\top \boldsymbol{\beta}\right) \boldsymbol{x}_j^\top \left\{\sum_{k\in s} exp\left(\boldsymbol{x}_k^\top \boldsymbol{\beta}\right) \boldsymbol{x}_k \boldsymbol{x}_k^\top\right\}^{-1} \boldsymbol{x}_i \left\{y_i - exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right\} & \text{if } i \in s \\ -\left\{y_i - exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right\} & \text{if } i \in U\setminus s \end{cases}$$

6.1.3 Construction of the robust predictor

Proposition 6.1. *The prediction error of* $\hat{\theta}^{EBP}$, $\hat{\theta}^{EBP} - \theta$ *can be expressed as*

$$\hat{\theta}^{EBP} - \theta = \sum_{i \in U \setminus s} B_i^{EBP}\left(I_i = 0\right) + \sum_{i \in s} B_i^{EBP}\left(I_i = 1\right) + o_p\left(\frac{1}{n^{1/2}}\right). \qquad (6.1.9)$$

The proof of the decomposition is given in the Appendix.

Remark 6.1. In the linear case, Beaumont et al. (2013) show that the prediction error of the BLUP $\hat{\theta}^{BLUP} - \theta$ can be exactly decomposed as

$$\hat{\theta}^{BLUP} - \theta = \sum_{i \in U \setminus s} B_i^{BLUP}\left(I_i = 0\right) + \sum_{i \in s} B_i^{BLUP}\left(I_i = 1\right).$$

The conditional bias can be viewed as the contribution of the unit i to the prediction error of the EBP. Nevertheless, nothing can be done at the estimation stage for the non-sampled units since the conditional bias of a non-sampled unit depends on the unknown y-value of this unit, which is by definition not observed. Robustness to influential units can be achieved by downweighting the contribution of the influential sampled units in (6.1.9). Following Beaumont et al. (2013), a robust version of the EBP is given by

$$\hat{\theta}^{REBP}\left(c\right) = \hat{\theta}^{EBP} - \sum_{i \in s} B_i^{EBP}\left(I_i = 1\right) + \sum_{i \in s} \psi_c\left\{B_i^{EBP}\left(I_i = 1\right)\right\}, \qquad (6.1.10)$$

where $\psi_c(.)$ is the Huber function, defined by $\psi_c(z) = sign(z) \times \min\left(|z|, c\right)$, where c is a positive tuning constant.

The conditional bias $B_i^{EBP}\left(I_i = 1\right)$ in (6.1.10) depends on unknown super-population parameters that should be estimated using the sample data. The estimation of $\boldsymbol{\beta}$ can be carried out by solving the sample estimating equation (6.1.4) using a Newton-Raphson algorithm. Replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ in the expression

(6.1.8) leads to the resulting estimation of the conditional bias $B_i^{EBP}(I_i=1)$:

$$\hat{B}_i^{EBP}(I_i=1) = \sum_{j\in U\setminus s} h_j^{'}(\hat{\eta}_j)\boldsymbol{x}_j^{\top}\left\{\sum_{k\in s}\boldsymbol{H}\left(\hat{\boldsymbol{\beta}}\right)\right\}^{-1}\boldsymbol{x}_i\hat{w}_i g_i^{'}(\hat{\mu}_i)\left\{y_i - h\left(\boldsymbol{x}_i^{\top}\hat{\boldsymbol{\beta}}\right)\right\}. \tag{6.1.11}$$

where $\hat{\eta}_j = \boldsymbol{x}_j^{\top}\hat{\boldsymbol{\beta}}$, $\hat{\mu}_i = h\left(\hat{\eta}_i\right) = h\left(\boldsymbol{x}_i^{\top}\hat{\boldsymbol{\beta}}\right)$ and $\hat{w}_i = \left\{v(\hat{\mu}_i)g_i^{'}(\hat{\mu}_i)\right\}^{-1}$.

Then, replacing $B_i^{EBP}(I_i=1)$ by $\hat{B}_i^{EBP}(I_i=1)$ in (6.1.10) leads to the following robust predictor

$$\hat{\theta}^{REBP}(c) = \hat{\theta}^{EBP} - \sum_{i\in s}\hat{B}_i^{EBP}(I_i=1) + \sum_{i\in s}\psi_c\left\{\hat{B}_i^{EBP}(I_i=1)\right\} \tag{6.1.12}$$

The robust predictor $\hat{\theta}^{REBP}$ depends on a tuning constant, which should be able to make the trade-off between bias and variance. Note that if the underlying tuning constant c in the Huber ψ function is large, the predictor $\hat{\theta}^{REBP}$ tends to the EBP $\hat{\theta}^{EBP}$. A suitable way to determine the tuning constant c is to find the tuning constant which minimizes an estimator of the mean square error. In this case, the mean square error expression in untractable. To cope with the problem, we consider an alternative criterion which consists on finding the value of c that minimizes

$$\max\left\{\left|\hat{B}_i^{REBP}(I_i=1)\right|, i=1,..,n\right\},$$

where $\hat{B}_i^{REBP}(I_i=1)$ is an estimator of the conditional bias attached to unit i of the robust Empirical Best Predictor. Using the definition of the conditional bias (6.1.6) applied to the robust predictor $\hat{\theta}^{REBP}(c)$, we have

$$\begin{aligned} B_i^{REBP}(I_i=1) &= E_m\left(\hat{\theta}^{REBP}(c) - \theta | s, Y_i = y_i\right) \\ &= B_i^{EBP}(I_i=1) + E_m\left\{\bar{\Delta}(c) | s, Y_i = y_i\right\}, \end{aligned} \tag{6.1.13}$$

where

$$\bar{\Delta}(c) = \sum_{i \in S} \left[\psi \left\{ B_i^{EBP} \left(I_i = 1 \right) \right\} - B_i^{EBP} \left(I_i = 1 \right) \right].$$

Using the estimation of the conditional bias for the EBP, $\hat{B}_i^{EBP} \left(I_i = 1 \right)$ and noting that $\bar{\Delta}(c)$ is a conditionally unbiased estimator of $E_m \left\{ \bar{\Delta}(c) | s, Y_i = y_i \right\}$, we have the following estimation for the conditional bias of the robust predictor $\hat{\theta}^{REBP}$

$$\hat{B}_i^{REBP} \left(I_i = 1 \right) = \hat{B}_i^{EBP} \left(I_i = 1 \right) + \sum_{i \in S} \left[\psi \left\{ \hat{B}_i^{EBP} \left(I_i = 1 \right) \right\} - \left(I_i = 1 \right) \right]. \quad (6.1.14)$$

Let $\hat{B}_{min} = \min \left\{ \hat{B}_i^{EBP} \left(I_i = 1 \right), \, i = 1, .., n \right\}$ and $\hat{B}_{max} = \max \left\{ \hat{B}_i^{EBP} \left(I_i = 1 \right), \, i = 1, .., n \right\}$. It is easily shown that the value of c that minimizes $\max \left\{ \left| \hat{B}_i^{REBP} \left(I_i = 1 \right) \right|, \, i = 1, .., n \right\}$ denoted by c_{minmax} , leads to the robust predictor

$$\hat{\theta}^{REBP} \left(c_{minmax} \right) = \hat{\theta}^{EBP} - \frac{1}{2} (\hat{B}_{min} + \hat{B}_{max}). \quad (6.1.15)$$

6.1.4 Simulation study

We conducted a simulation study in order to assess the empirical bias and mean square error of the several predictors. First, we described the mixture model.

Let A_i be Bernoulli random variables with parameter p, which represents the proportion of outliers. We used $p = 0;\ 0.005;\ 0.01$ and $p = 0.05$.

For the logistic model, we considered the mixture model $\xi_{\mathcal{B}}$ where

$$y_i = A_i\mathcal{B}(m, p_i) + (1 - A_i)\mathcal{B}(m, 0.8)$$

with $m = 20$, $p_i = \exp(0.2 \times x_i - 5)/\left\{1 + \exp\left(0.2 \times x_i - 5\right)\right\}$ and

$$x_i = A_i\mathcal{N}\left(10, 10\right) + (1 - A_i)\mathcal{N}\left(-20, 1\right).$$

For the Poisson model, we considered the mixture model $\zeta_{\mathcal{P}}$:

$$y_i = A_i\mathcal{P}(\eta_i) + (1 - A_i)\mathcal{P}(15)$$

with $\eta_i = \exp\left\{0.5\left(x - 250\right)/70\right\}$ and

$$x_i = A_i\mathcal{N}\left(350, 70\right) + (1 - A_i)\mathcal{N}\left(270, 30\right).$$

Figure 6.1.1 represent one realisation of the population under the two models $\xi_{\mathcal{B}}$ and $\xi_{\mathcal{P}}$. The population 1 and 3 represented in Figure 6.1.1a and 6.1.1c are generated without outliers, whereas the populations 2 and 4 represented in Figure 6.1.1b and 6.1.1d were generated with an outlier rate $p = 0.05$.

We conducted $P = 5000$ Monte-Carlo replications. For each replication, a population was drawn according to the mixture model $\xi_{\mathcal{B}}$ or $\zeta_{\mathcal{P}}$, then a random sample of size $n = 100; 300; 500$ was selected by simple random sampling without remplacement and we computed the four predictor: (i) the Empirical Best Predic-

tor $\hat{\theta}^{EBP}$ defined by expression (6.1.3) (ii) our robust predictor $\hat{\theta}^{REBP}$ defined by expression (6.1.15) (iii) the robust predictor $\hat{\theta}^{RCantoni}$ (6.1.5) using a M-estimator and the tuning constant $c_R = 1.345$ (iv) an oracle version $\hat{\theta}^{RCantoni}(c_{opt})$ of the robust predictor (6.1.5), where the tuning constant $c_R = c_{opt}$ minimizes the Monte-Carlo estimated mean square error. This estimator is impossible to compute in practice, but it will be usefull for comparison . We limited our empirical study to logistic and Poisson cases and considered one mixture model for each case.

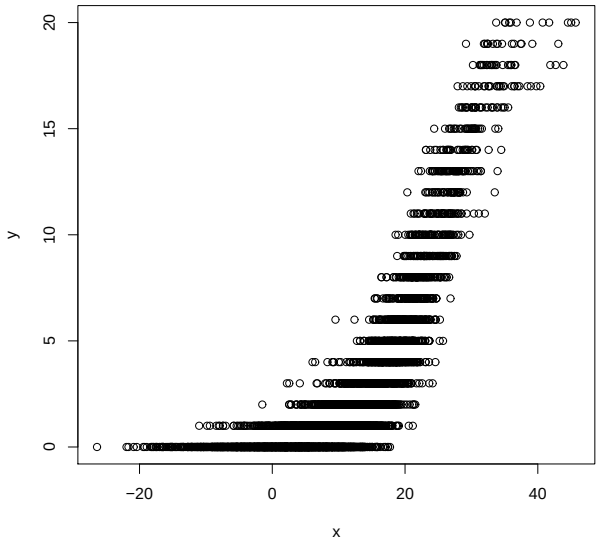
Let $\hat{\theta}$ be a generic estimator of θ. As a measure of bias, we computed the Monte Carlo percent relative bias (RB) given by

$$RB_{MC}\left(\hat{\theta}\right) = \frac{1}{P}\sum_{p=1}^{P}\frac{\left(\hat{\theta}_{(p)} - \theta_{(p)}\right)}{\theta_{(p)}} \times 100,$$

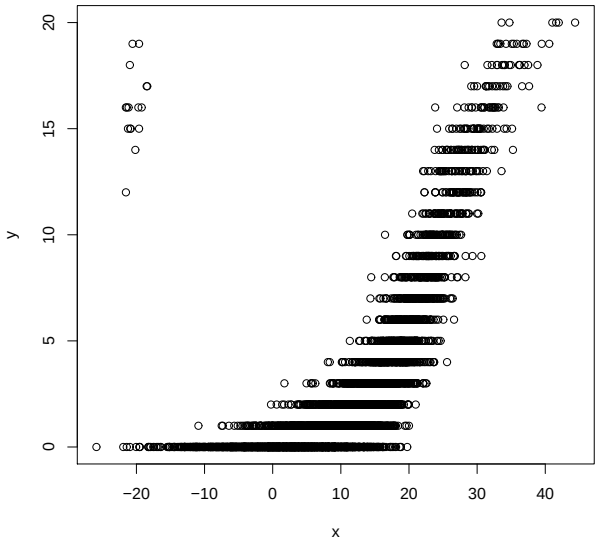
with $\hat{\theta}_{(p)}$ denoting the estimator $\hat{\theta}$ in the p-th iteration, $p = 1, \ldots, 5,000$.In matrix notation, As a measure of efficiency, using the Empirical Best Predictor $\hat{\theta}^{EBP}$ as the reference, we computed the Monte Carlo Relative Efficiency(RE)

$$RE_{MC}\left(\hat{\theta}, \hat{\theta}^{EBP}\right) = \frac{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}_{(p)} - \theta_{(p)}\right)^2}{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}_{(p)}^{EBP} - \theta_{(p)}\right)^2} \times 100.$$

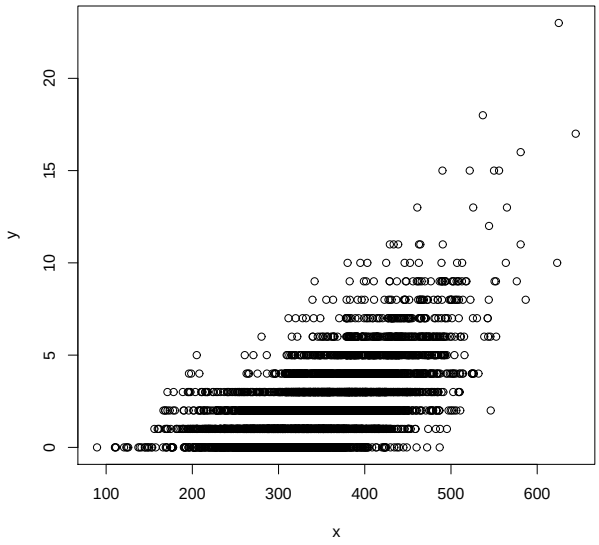
where $\hat{\theta}_{(p)}^{EBP}$ denotes the EBP in the p-th iteration, $p = 1, \ldots, 5,000$.



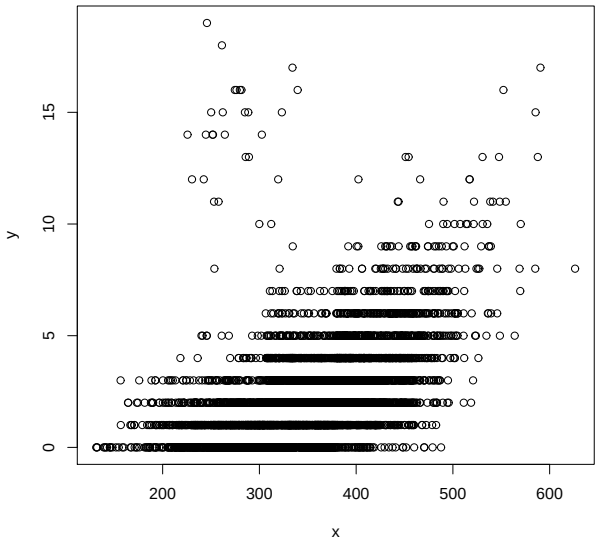
(a) Population 1



(b) Population 2



(c) Population 3



(d) Population 4

Figure 6.1.1: Representation of the four populations for p=0.05

Population	Outlier rate p	Sample size	$\hat{\theta}^{REBP}$	$\hat{\theta}^{RCantoni}$	$\hat{\theta}^{RCantoni}(c_{opt})$
Binomial	0	50	−0.0(100)	0.1(105)	0.1(100)
		100	−0.1(100)	−0.0(105)	−0.0(100)
		500	−0.0(100)	0.0(105)	0.0(100)
	0.5%	50	−1.7(80)	−2.8(59)	−2.9(58)
		100	−1.2(82)	−2.9(72)	−2.8(68)
		500	−0.5(97)	−2.7(153)	−0.2(96)
	1%	50	−3.0(82)	−5.7(55)	−5.8(54)
		100	−2.0(87)	−5.8(82)	−5.6(77)
		500	−0.5(98)	−5.2(261)	−0.7(97)
	5%	50	−6.0(103)	−24.5(180)	−2.3(100)
		100	−3.3(104)	−23.9(311)	−1.2(100)
		500	−0.7(101)	−21.9(1367)	−0.3(100)
Poisson	0	50	−0.1(100)	−0.3(105)	−0.2(100)
		100	−0.2(100)	−0.3(106)	−0.0(100)
		500	−0.1(100)	−0.3(107)	−0.0(100)
	0.5%	50	−1.5(88)	−2.6(82)	−2.2(78)
		100	−1.1(90)	−2.7(88)	−2.0(81)
		500	−0.38(97)	−2.5(138)	−0.7(93)
	1%	50	−2.3(86)	−5.0(79)	−4.0(74)
		100	−1.6(90)	−4.9(94)	−3.1(79)
		500	−0.5(99)	−4.7(221)	−0.8(97)
	5%	50	−4.1(99)	−19.5(166)	−0.5(100)
		100	−2.4(100)	−19.5(304)	−0.2(100)
		500	−0.6(100)	−18.0(1351)	−0.0(100)

Table 6.1.1: Bias and relative efficiency in brackets of the robust predictors

The results are shown in Table 6.1.1. Under the logistic and Poisson models (cases where $p = 0$), we note that the bias of the three robust predictors were small (less than 0.1%) for any sample size n. Furthermore, the robust predictor $\hat{\theta}^{REBP}$ was nearly as efficient as the Empirical Best Predictor and we observe a sight loss of efficiency for the predictor $\hat{\theta}^{RCantoni}$, which is consistent with the property of M-estimators; they have a 95% asymptotic efficiency with the tuning constant $c = 1.345$. Note that in the case where the outlier rate p is equal to 5%, there were no significant improvements for any sample size n for $\hat{\theta}^{REBP}$, but the loss of efficiency was substantial for $\hat{\theta}^{RCantoni}$ with value of efficiency ranging

from 166% to 1367% . In the case where the outlier rate is equal to either 1% or 0.5% and the sample size is equal to either 50 or 100, the two robust predictors perform better than the EBP with values of the relative efficiency ranging from 80% to 90% for $\hat{\theta}^{REBP}$ and from 55% to 90% for $\hat{\theta}^{RCantoni}$. Note that in these cases, the relative bias is under 6%, which seems reasonable. Finally, for $n = 500$ and $p = 0.01$ or $p = 0.005$, we can see that there were no significant improvements for $\hat{\theta}^{REBP}$ compare to the $\hat{\theta}^{EBP}$, but we observe a significant loss of efficiency for $\hat{\theta}^{RCantoni}$ with values ranging from 138% to 261%.

6.2 Robust Small area prediction using GLMMs

6.2.1 Small area prediction based on GLMMs

We now turn our attention on the Generalized Linear Mixed Model (GLMM) with application to small area estimation. In a small−area context, the empirical best linear unbiased predictors or plug-in predictors are efficient under correct model specification and distributional assumptions, but they may be highly sensitive to presence of outliers. In the case of LMMs, Gosh et al. (2008), Sinha and Rao (2009), Dongmo Jiongo et al. (2013) and Chambers et al. (2014) show that the presence of influential units in the sample tends to make empirical best linear predictor unstable. The main objective here is to propose new robust predictors for small areas means using the conditional bias. We propose an extension of the work of Dongmo Jiongo et al. (2013) to GLMMs and consider, following Beaumont et al. (2013), an adaptive choice of the tuning constant by using a minmax criterion, i.e we choose the tuning constant which minimizes the maximum of the absolute value of the conditional bias of the robust predictor.

To extend the results in the case of small area prediction, we slightly modify the notation introduced in the previous sections. Let U denote the finite population

of size N, which is partitioned into D domains (or small areas) $U_1, .., U_D$ of size $N_1, ...N_D$, respectively. The domains sizes N_d are assumed to be known. Let y_{dj} be the value of $y-$variable attached to unit j in area d and let $\boldsymbol{x}_{dj}$ be a deterministic $p-$vector containing the unit level covariates for unit j in the area d. It is assumed that the values of $\boldsymbol{x}_{dj}$ are recorded without error for all units in the population. A sample s of size n is selected from U according to a non-informative sampling plan $p(s)$. Let $s_d = s \cap U_d$ denote the sample of size n_d. The aim is to use the sample values of y_{dj} and the populatIn matrix notation, ion values $\boldsymbol{x}_{dj}$ to infer on the small area means $\theta_d = N_d^{-1} \sum_{j \in U_d} y_{dj}$. Let $\boldsymbol{y} = (y_{1,1}.., y_{1,N_1}, .., y_{D,1}, .., y_{D,N_D})^\top$be the $N-$vector of the $y-$values and $\boldsymbol{u} = (u_1, ..., u_D)^\top$ be the $D-$vector of random area effects.Let $\boldsymbol{\mu} = E_m(\boldsymbol{y}|\boldsymbol{u})$ be the conditional mean vector with elements μ_{dj} and $\boldsymbol{\Sigma} = Var_m(\boldsymbol{y}|\boldsymbol{u})$ be the conditional covariance matrix which is diagonal with element σ_{dj}. Let us define the $N \times D$ matrix $\boldsymbol{Z} = diag(1_{N_d}, d = 1, .., D)$ where $\mathbf{1}_{N_d}$denotes the N_d-vector of ones .

We assume a generalized linear mixed model (GLMM) for μ_{dj} of the form

$$g(\mu_{dj}) = \eta_{dj} = \boldsymbol{x}_{dj}^\top \boldsymbol{\beta} + u_d, \tag{6.2.1}$$

where $g(.)$ is the link function, assumed to be known and invertible.

Assuming that the model (6.2.1) holds, the minimum mean square error predictor of θ_d is

$$\frac{1}{N_d}\left(\sum_{j \in s_d} y_{dj} + \sum_{j \in U_d \setminus s_d} \mu_{dj}\right).$$

Since μ_{dj} depends on $\boldsymbol{\beta}$ and u_d, we need to provide an estimation of $\boldsymbol{\beta}$ and u_d ,

and this leads to a Empirical Plug-in Predictor of the $d-th$ area mean,

$$\hat{\theta}_d^{EPP} = \frac{1}{N_d}\left(\sum_{j\in s_d} y_{dj} + \sum_{j\in U_d\setminus s_d} \hat{\mu}_{dj}\right) \tag{6.2.2}$$

where $\hat{\mu}_{dj} = h\left(\boldsymbol{x}_{dj}^{\top}\hat{\boldsymbol{\beta}} + \hat{u}_d\right)$, $\hat{\boldsymbol{\beta}}$ is the vector of the estimated fixed effect, $\hat{u}_d$ is the predicted area random effect for the area d and $h(.)$ is the inverse of the link function $g(.)$.

The EPP is efficient under correct model specifications and distributional assumptions. More details on this predictor, including MSE estimation are given in Saei & Chambers (2003), Jiang & Lahiri (2006) and González-Manteiga et al. (2007). Despite its attractive properties, the EPP might be very sensitive to the presence of outliers, especially when the sample size in small. The main objective here is to propose new robust predictors for small-area means using the conditional bias as a measure of influence.

6.2.2 Conditional bias of the EPP based on GLMMs

Since the EPP is no longer linear on the $y-$values, the conditional bias can't be computed analytically. That's why we consider a linear mixed model which approximates the original generalized mixed model and then develop an approximation of the conditionnal bias under the approximation version of the model.

Following González-Manteiga et al. (2007), a first-order Taylor expansion of $g(y_{dj})$ around μ_{dj}, ignoring the higher-order terms, leads to

$$g(y_{dj}) \simeq \eta_{dj} + (y_{dj} - \mu_{dj})\, g^{'}(\mu_{dj}) \triangleq \xi_{dj}. \tag{6.2.3}$$

The conditional moments of the working variables ξ_{dj} are given by

$$E_m\left(\xi_{dj}|\boldsymbol{u}\right)=\eta_{dj},\quad Var\left(\xi_{dj}|\boldsymbol{u}\right)=g^{'}\left(\mu_{dj}\right)^2\sigma_{dj}^2$$

and

$$Cov_m\left(\xi_{dj},\xi_{d^{'}j^{'}}|\boldsymbol{u}\right)=0,\ \text{for}\ d\neq d^{'}\ \text{or}\ j\neq j^{'}.$$

The unconditional mean of ξ_{dj} is $\boldsymbol{x}_{dj}^{\top}\boldsymbol{\beta}$, and the random effects u_d are assumed to be independent, normally distributed, with zero mean and constant variances equal to σ_u.

Now, we approximate the original generalized mixed model (6.2.3) by the linear mixed model

$$\xi_{dj}=\boldsymbol{x}_{dj}^{\top}\boldsymbol{\beta}+u_d+e_{dj},\ j=1,...,N_i,\ d=1,...,D,$$

where e_{dj} are independent random variables, independent of $\boldsymbol{u}$, with zero means and variance $v_{dj}=g^{'}\left(\mu_{dj}\right)^2\sigma_{dj}^2$ and

$$Var_m\left(\xi_{dj}\right)=\sigma_u\boldsymbol{Z}_s\boldsymbol{Z}_s^{\top}+\boldsymbol{\Sigma}_{es}\triangleq\boldsymbol{V}_s,$$

where $\boldsymbol{\Sigma}_{es}$ is a diagonal matrix whose elements are the variances v_{dj} of the residuals e_{dj}.

In matrix notation, this model can be compactly rewritten as

$$\boldsymbol{\xi}_d=\boldsymbol{X}_d\boldsymbol{\beta}+u_d\mathbf{1}_{n_d}+\boldsymbol{e}_i\ \ (d=1,...,D),$$

where $\boldsymbol{X}_d=(\boldsymbol{x}_{d1},..,\boldsymbol{x}_{dn_d})$ is a matrix of dimension $n_d\times p$ and $\mathbf{1}_{n_d}$ corresponds to a n_d-vector of ones. The variance-covariance matrix of $\boldsymbol{\xi}_d$ is $\boldsymbol{V}_d=\sigma_u^2\mathbf{1}_{n_d}\mathbf{1}_{n_d}^{\top}+\boldsymbol{\Sigma}_{es_i}$ where $\boldsymbol{\Sigma}_{es_d}$ correspond to the $d-th$ block of the matrix $\boldsymbol{\Sigma}_{es}$.

We assume that the variance matrix $\boldsymbol{V}_s$ is known. In this case, the best linear unbiased estimator of $\boldsymbol{\beta}$ and the best linear predictor of $\boldsymbol{u}$ in the linear mixed model are given by

$$\hat{\boldsymbol{\beta}} = \left(\sum_{h=1}^{D} \boldsymbol{X}_h^\top \boldsymbol{V}_h^{-1} \boldsymbol{X}_h \right)^{-1} \sum_{h=1}^{D} \boldsymbol{X}_h^\top \boldsymbol{V}_h^{-1} \boldsymbol{\xi}_h$$

and

$$\hat{u}_d = \sigma_u^2 \mathbf{1}_{n_h}^\top \boldsymbol{V}_h^{-1} \left(\boldsymbol{\xi}_h - \boldsymbol{X}_h \hat{\boldsymbol{\beta}} \right).$$

Note that the approximation of the conditional bias of the EPP, will not take into account the impact of a unit on the estimation of the variance components appearing in the EPP, since we assume that the variance matrix $\boldsymbol{V}_s$ is known.

As a measure of influence of a unit, we consider the concept of conditional bias. Following Dongmo Jiongo et al. (2013), we compute the conditional bias of a unit j in the area h for the Empirical Plug-in Predictor of the d-th area , but in order to simplify the derivations, we use a conditioning on all area-random effect instead of the random effect in the d-th area.

$$B_{dhj}\left(y_{hj}, u_h; \boldsymbol{\beta}\right) \;=\; E_m\left(\hat{\theta}_d^{EPP} - \theta_d | s,\, y_{hj},\, \boldsymbol{u} \right). \tag{6.2.4}$$

In other words, we calculate the conditional expectation, keeping fixed the y-value of the unit under consideration and also all the local area effects $\boldsymbol{u}$. Consequently, the conditional bias measures the average effect of the unit j in area h, on the predictor $\hat{\theta}_d^{EPP}$. To determine this conditional bias, we need to distinguish four cases, whether the unit j belongs to the domain d or not and whether the unit j is sampled or not and we have to keep in mind that the weights w_{dhj} depends on all area random effects $\boldsymbol{u}$.

The conditional bias of $\hat{\theta}_d^{EPP}$ attached to the unit j in domain h is approxi-

matively given by

$$B_{dhj}\left(y_{hj}, u_h; \boldsymbol{\beta}\right) \simeq \begin{cases} N_d^{-1}\left(\sum_{h=1}^{D}\sum_{j\in s_h} w_{dhj}u_h - \sum_{j\in U_d\setminus s_d} h^{'}\left(\eta_{dj}\right)u_d + w_{ddj}e_{dj}\right) & j\in s_d \\ N_d^{-1}\left(\sum_{h=1}^{D}\sum_{j\in s_h} w_{dhj}u_h - \sum_{j\in U_d\setminus s_d} h^{'}\left(\eta_{dj}\right)u_d + w_{dhj}e_{hj}\right) & j\in s_h, h\neq d \\ N_d^{-1}\left(\sum_{h=1}^{D}\sum_{j\in s_h} w_{dhj}u_h - \sum_{j\in U_d\setminus s_d} h^{'}\left(\eta_{dj}\right)u_d - \frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)e_{dj}\right) & j\in U_d\setminus s_d \\ N_d^{-1}\left(\sum_{h=1}^{D}\sum_{j\in s_h} w_{dhj}u_h - \sum_{j\in U_d\setminus s_d} h^{'}\left(\eta_{dj}\right)u_d\right) & j\in U_h\setminus s_h, h\neq d, \end{cases} \tag{6.2.5}$$

where

$$w_{dhj} = \begin{cases} D^{-1}a_d\boldsymbol{X}_h^{\top}\boldsymbol{C}_h^{(j)} & j\in s_h \\ D^{-1}a_d\boldsymbol{X}_d^{\top}\boldsymbol{C}_d^{(j)} + \left[\sum_{l\in U_d\setminus s_d} h^{'}\left(\eta_{dl}\right)\right]\sigma_u^2\mathbf{1}_{n_d}^{\top}\boldsymbol{C}_d^{(j)} & j\in s_d \end{cases}$$

$$a_d = \left\{\sum_{j\in U_d\setminus s_d}\frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)\left(\boldsymbol{x}_{dj}^{\top} - \sigma_u^2\mathbf{1}_{n_d}^{\top}\boldsymbol{V}_d^{-1}\boldsymbol{X}_d\right)\right\}\left(D^{-1}\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\boldsymbol{X}_h\right)^{-1}$$

and

$$e_{dj} = \left(y_{dj} - \mu_{dj}\right)g^{'}\left(\mu_{dj}\right).$$

The details are given in the Appendix.

We note that a unit outside the area $j\in s_h$ with $h\neq d$ may have a large influence if its weight w_{dhj} is large and its model residual e_{hj} is large. Also even if a non-sampled unit has a large influence, it is not possible to reduce its impact at the estimation stage, because its conditional bias can not be estimated from the sample.

Now, it can be shown that the prediction error of the EPP $\hat{\theta}_d^{EPP} - \theta_d$ can be written as

$$\hat{\theta}_d^{EPP} - \theta_d = \sum_{h=1}^{D} \sum_{j \in U_h} B_{dhj}\left(y_{hj}, u_h; \beta\right) - \frac{N-1}{N_d}\left[\sum_{h=1}^{D} \sum_{j \in s_h} w_{dhj} u_h - \sum_{j \in U_d \setminus s_d} h^{'}\left(\eta_{dj}\right) u_d\right]. \tag{6.2.6}$$

We note from (6.2.6) that a unit exhibiting a large conditional bias will contribute in increasing the prediction error of $\hat{\theta}_d^{EPP}$.

6.2.3 Construction of the robust predictor

Following Dongmo Jiongo et al. (2014), we define a robust predictor of θ_d as

$$\hat{\theta}_d^{REPP} = \hat{\theta}_d^{EPP} - \sum_{h=1}^{D} \sum_{j \in s_h} B_{dhj}^{'}\left(y_{hj}, u_h; \boldsymbol{\beta}\right) + \sum_{h=1}^{D} \sum_{j \in s_h} \psi_{d_1}\left\{B_{dhj}^{'}\left(y_{hj}, u_h; \boldsymbol{\beta}\right)\right\} \tag{6.2.7}$$

where

$$B_{dhj}^{'}\left(y_{hj}, u_h; \boldsymbol{\beta}\right) = \begin{cases} N_d^{-1} w_{ddj} e_{dj} & j \in s_d \\ N_d^{-1} w_{dhj} e_{hj} & j \in s_h, h \neq d \end{cases}$$

if the influence of all the sample units is small, we have

$$\psi_{d_1}\left\{B_{dhj}^{'}\left(y_{hj}, u_h; \boldsymbol{\beta}\right)\right\} = B_{dhj}^{'}\left(y_{hj}, u_h; \boldsymbol{\beta}\right), \forall j \in s,$$

so the summation of the second and third term is close to zero, therefore the robust predictor is close to the non-robust one, i.e the EPP.

The conditional bias $B_{dhj}^{'}\left(y_{hj}, u_h; \boldsymbol{\beta}\right)$ is unknown since they depend on the model parameters ($\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{es}$), the random small-area effects $\boldsymbol{u}$, and the variance of the small area effects σ_u^2. The estimation of these parameters can be carried out by using a combination of Maximum Penalized Quasi-Likelihood (MPQL) for the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{u}$, and REML for the estimation of the variance components (Schall, 1991; Saei and Chambers, 2003). We denote by $\hat{e}_{dj}$ and $\hat{w}_{dhj}$ the plug-

in estimators of e_{dj} and w_{dhj}. By plugging these estimators into the expression (6.2.5) of the conditional bias $B'_{dhj}\left(y_{hj}, u_h; \boldsymbol{\beta}\right)$, this leads to a plug-in estimator denoted by

$$\hat{B}'_{dhj}\left(y_{hj}, \hat{u}_h; \hat{\boldsymbol{\beta}}\right) = \begin{cases} N_d^{-1}\hat{w}_{ddj}\hat{e}_{dj} & j \in s_d \\ N_d^{-1}\hat{w}_{dhj}\hat{e}_{hj} & j \in s_h, h \neq d \end{cases}$$

Then the robust predictor (6.2.7) using a sample approximation of the conditional bias becomes

$$\hat{\theta}_d^{REPP} = \hat{\theta}_d^{EPP} - \sum_{h=1}^{D}\sum_{j\in s_h} \hat{B}'_{dhj}\left(y_{hj}, \hat{u}_h; \hat{\boldsymbol{\beta}}\right) + \sum_{h=1}^{D}\sum_{j\in s_h} \psi_{d_1}\left\{\hat{B}'_{dhj}\left(y_{hj}, \hat{u}_h; \hat{\boldsymbol{\beta}}\right)\right\}.$$

We have to determine the tuning constant d_1 which adjust the trade-off between bias and variance. Since, it is not possible to provide an analytic expression for the mean square error of the robust predictor, we choose the constant d_1 which verify a minmax criterion. We can use the same proof as in Section 6.1 to show that the robust predictor construct with the minmax constant can be written as

$$\hat{\theta}_d^{REPP} = \hat{\theta}_d^{EPP} - \frac{1}{2}\left(\hat{B}'_{max} + \hat{B}'_{min}\right) \tag{6.2.8}$$

where $\hat{B}'_{max} = \max_{j\in s}\left\{\hat{B}'_{dhj}\left(y_{hj}, \hat{u}_h; \hat{\boldsymbol{\beta}}\right)\right\}$ and $\hat{B}'_{min} = \min_{j\in s}\left\{\hat{B}'_{dhj}\left(y_{hj}, \hat{u}_h; \hat{\boldsymbol{\beta}}\right)\right\}$.

6.2.4 Simulation studies

6.2.4.1 Linear case

We conducted a Monte Carlo study in order to assess the empirical bias and mean square error of several predictors. We considered the mixture model ζ_m of two

unit-level models :

$$\begin{aligned}\zeta_0 : y_{0dj} &= \beta_{00} + \beta_{01}x_{dj} + u_{0d} + e_{0dj} \qquad (j = 1, ..., N_d; d = 1, ..., D),\\ \zeta_1 : y_{1dj} &= \beta_{10} + \beta_{11}x_{dj} + u_{1d} + e_{1dj} \qquad (j = 1, ..., N_d; d = 1, ..., D).\end{aligned}$$

We considered $D = 40$ domains, each of size $N_d = 50$, $d \in [\![1 : 40]\!]$. The values of the auxiliary information were generated from a normal distribution with $E(X) = 2$ and $\sqrt{Var(X)} = 0.35$. The mixture model ζ_m satisfied $y_{dj} = (1 - A_{dj})\, y_{0dj} + A_{dj} y_{1dj}$, where the A_{dj} are generated according to a Bernoulli distribution, with parameter p. The values of p were set to 0.01; 0.05, and 0.01. In each area, random samples of size $n_d = 5$, $d \in [\![1 : 40]\!]$ were selected according to simple random sampling without remplacement. The choice of the variance and slopes parameter is detailed in Table 6.2.1.

Scenarios	Variances error terms	Variances random effects	Intercepts and Slopes	Distribution Error
(0,0,0)	$(\sigma^2_{e0}, \sigma^2_{e1}) = (6, 6)$	$(\sigma^2_{u0}, \sigma^2_{u1}) = (6, 6)$	$(\beta_{00}, \beta_{01}) = (\beta_{10}, \beta_{11}) = (100, 3)^\top$	Normal
(e,u,0)	$(\sigma^2_{e0}, \sigma^2_{e1}) = (6, 150)$	$(\sigma^2_{u0}, \sigma^2_{u1}) = (6, 150)$	$(\beta_{00}, \beta_{01}) = (\beta_{10}, \beta_{11}) = (100, 3)^\top$	Normal
(e,u,b)	$(\sigma^2_{e0}, \sigma^2_{e1}) = (6, 150)$	$(\sigma^2_{u0}, \sigma^2_{u1}) = (6, 150)$	$(\beta_{00}, \beta_{01}) = (100, 3)^\top, (\beta_{10}, \beta_{11}) = (150, 1)^\top$	Normal

Table 6.2.1: Sources of contamination

Five predictors of the small-area mean θ_d were included in the study. The empirical best linear unbiased predictor, based on the empirical best linear unbiased estimators and empirical best linear unbiased predictors, with variance components estimated by maximum likelihood. As in Sinha & Rao (2009), the robust predictors relied on the robustified maximum likelihood estimators of $(\boldsymbol{\beta}, \boldsymbol{V_s})$. The robust random effects were estimated by solving Fellner's equation. We used $b = 1 \cdot 345$. We computed also the robust predictor $\hat{\theta}_d^{CCST}$ proposed by Chambers et al. (2014) (refer as the REBLUP-BC predictor in their paper) with tuning constant set to $c = 3$. We also computed the two fully bias-corrected robust predictor $\hat{\theta}_d^{CB}$ and $\hat{\theta}_d^{C}$ defined respectively by the equation (16) and (25) in Dongmo

Jiongo et al. (2014) with the tuning constant set to $b = 1 \cdot 345$ or 3, 6, 9 . Finally, we computed our robust predictors $\hat{\theta}^{CminmaxR}$ and $\hat{\theta}^{Cminmax}$ using the minmax criterion with a robust and a non-robust estimation of the conditional bias, i.e the non-robust estimation of the conditional bias was obtained by replacing $(\boldsymbol{\beta}, \boldsymbol{V_s})$ by their maximum likelihood estimation and the robust estimation of the condtional bias was obtained by replacing $(\boldsymbol{\beta}, \boldsymbol{V_s})$ by their robustified maximum likelihood estimators and the robust random effects were estimated by solving Fellner's equation.

For each scenario described in Table 6.2.1, $I = 10000$ populations were generated. Let $\hat{\theta}_{d_{(i)}}$ denote a predictor for domain d at iteration i. The empirical percent absolute relative bias for the area mean θ_d was calculated as

$$ARB_{MC}\left(\hat{\theta}_d\right) = \left| \frac{1}{I} \sum_{i=1}^{I} \frac{\left(\hat{\theta}_{d_{(i)}} - \theta_{d_{(i)}}\right)}{\theta_{d_{(i)}}} \right| \times 100,$$

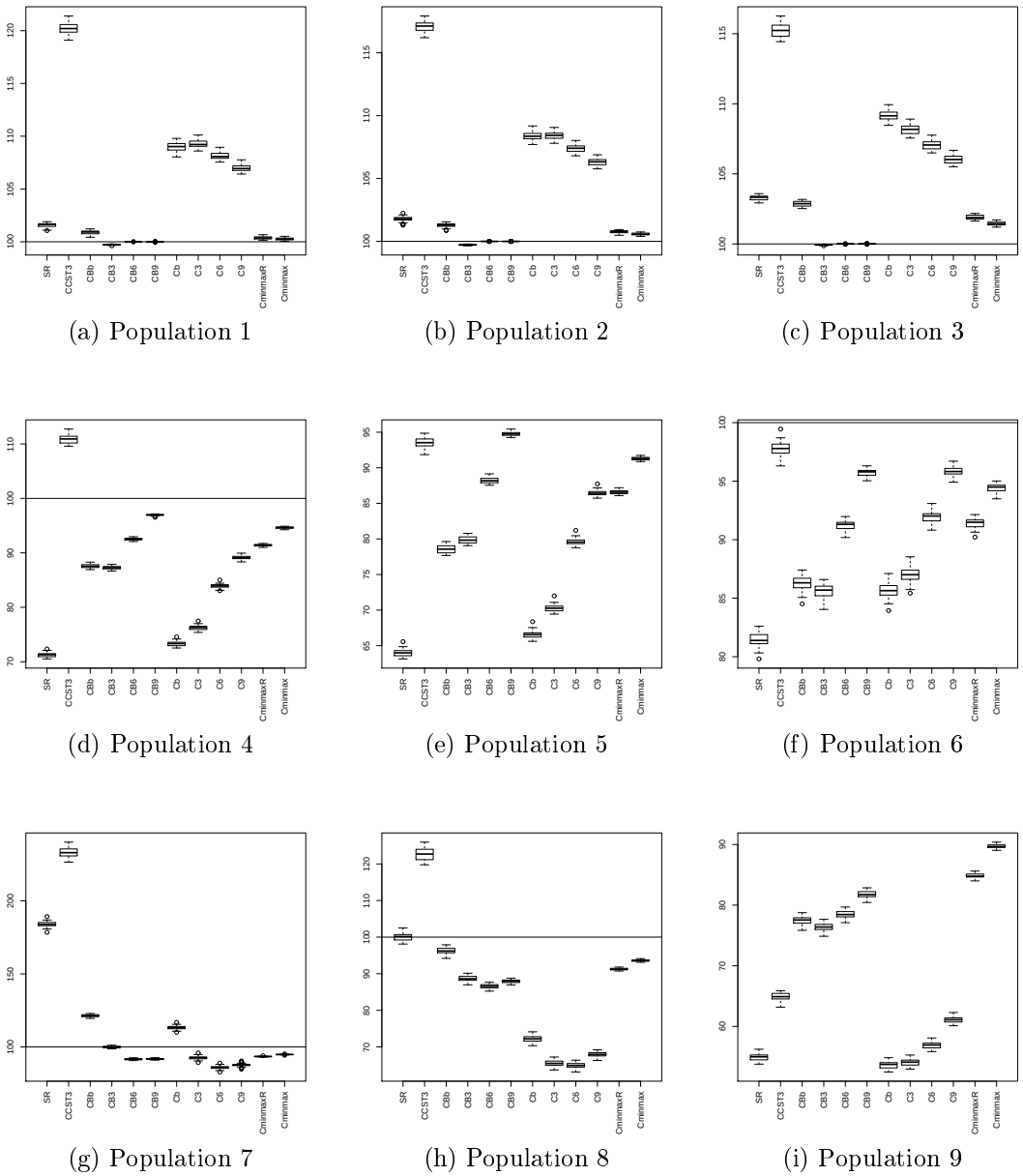
As a measure of efficiency, using the Empirical Best Predictor $\hat{\theta}^{EBLUP}$ as the reference, we computed the Monte Carlo Relative Efficiency(RE)

$$RE_{MC}\left(\hat{\theta}_d, \hat{\theta}_d^{EBLUP}\right) = \frac{\frac{1}{I}\sum_{i=1}^{I}\left(\hat{\theta}_{d_{(i)}} - \theta_{d_{(i)}}\right)^2}{\frac{1}{I}\sum_{i=1}^{I}\left(\hat{\theta}_{d_{(i)}}^{EBLUP} - \theta_{d_{(i)}}\right)^2} \times 100.$$

where $\hat{\theta}_{d_{(i)}}^{EBLUP}$ denotes the ELUP in the i-th iteration, $i = 1, \ldots, 10,000$.

Scenarios	N°	rate p	EBLUP	SR	CCST3	CBb	CB3	CB6	CB9	Cb	C3	C6	C9	CminmaxR	Cminmax
(0,0,0)	1	0.1	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01
	2	0.05	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01
	3	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01
(e,u,0)	4	0.1	0.00	0.02	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03
	5	0.05	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02
	6	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
(e,u,b)	7	0.1	0.00	3.13	2.37	1.72	1.05	0.55	0.30	1.68	1.01	0.52	0.28	0.27	0.27
	8	0.05	0.00	1.65	1.37	0.94	0.62	0.39	0.26	0.90	0.59	0.35	0.22	0.22	0.22
	9	0.01	0.00	0.33	0.29	0.17	0.14	0.10	0.07	0.15	0.12	0.08	0.06	0.06	0.06

Table 6.2.2: Average absolute relative bias for the predictors over areas



(a) Population 1 (b) Population 2 (c) Population 3

(d) Population 4 (e) Population 5 (f) Population 6

(g) Population 7 (h) Population 8 (i) Population 9

Figure 6.2.1: Populations

The bias averages over the areas were computed for each predictor and are shown in Table 6.2.2, where we see that all the methods exhibited small empirical biases for the scenarios $(0, 0, 0)$ and $(e, v, 0)$. Under the scenario (e, v, b) , we observe larger bias for all the robust predictors. We notice that the bias in a

increasing function of the rate of outliers and decreasing function of the tuning constant. Indeed, as expected, small values of the tuning constants generate larger empirical bias. From Figure 6.2.1, we see that under the scenario $(0, 0, 0)$, the Sinha-Rao predictor and the predictor based on the conditional bias perform as well as the empirical best linear unbiased predictor, so in absence of outlier, they are close to the EBLUP. Figure 6.2.1 shows that the case of the scenarios $(e, v, 0)$ and (e, v, b), the robust predictors based on the conditional bias with an adaptative tuning constant perform better than the EBLUP.

6.2.4.2 Poisson case

We conducted anothersimulation study, using this time, a mixture of two Poisson models. We consider the mixture model ζ_m :

$$\begin{aligned} \zeta_0 : y_{0dj} &\sim \mathcal{P}(\eta_{0dj}) \qquad (j = 1, ..., N_d; d = 1, ..., D), \\ \zeta_1 : y_{1dj} &\sim \mathcal{P}(\eta_{0dj}) \qquad (j = 1, ..., N_d; d = 1, ..., D), \end{aligned}$$

where

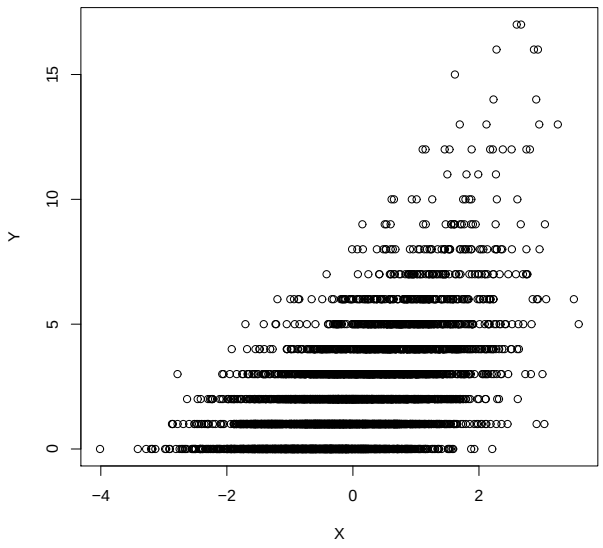
$$\begin{aligned} \eta_{0dj} &= \beta_{00} + \beta_{01} x_{dj} + u_{0d} \qquad (j = 1, ..., N_d; d = 1, ..., D), \\ \eta_{1dj} &= \beta_{10} + \beta_{11} x_{dj} + u_{1d} \qquad (j = 1, ..., N_d; d = 1, ..., D). \end{aligned}$$

We considered $D = 50$ domains of size $N_d = 100$, $d = 1, .., 50$. The values of the auxiliary information were generated from a normal distribution with $E(X) = 0$ and $Var(X) = 1$. The mixture model ζ_m satisfied $y_{dj} = (1 - A_{dj})\, y_{0dj} + A_{dj} y_{1dj}$, where the A_{dj} were generated according to a Bernoulli distribution, with different expectation $p = 0.1, p = 0.05$, or $p = 0.01$. In each area, random samples of size $n_d = 10, d \in [\![1 : 50]\!]$ are selected by simple random sampling without replacement. The choice of the variance and slopes parameter is detailed in Table 6.2.3.

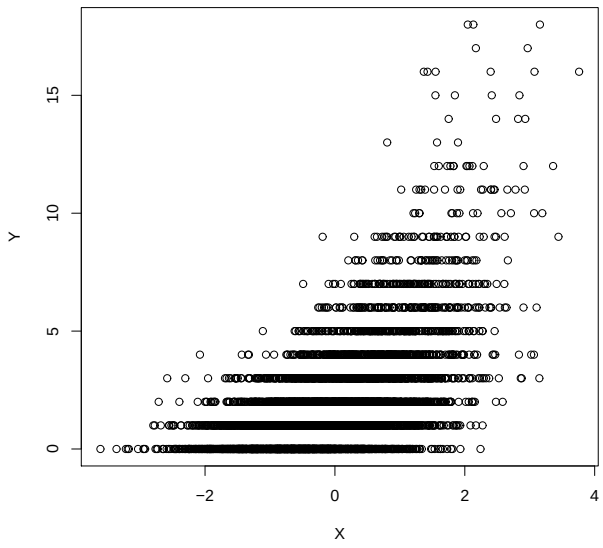
We also add a case with some measurement errors to the model ζ_0, we modify randomly some y-values, i.e $\tilde{y}_{dj} = y_{dj} + 10$ according to different rates $p = 0.1, p = 0.05$, or $p = 0.01$. Figure 6.2.2 shows one realisation of these models for $p = 0.01$. We compute five predictors: the EPP defined by expression (6.2.2), the robust M-quantile predictor (Tzavidis, 2014), our robust predictor defined by expression (6.2.8) , the synthetic predictor using a GLM without random effects for domains, and the direct estimator.

Scenarios	Variances random effects	Intercepts and Slopes
(0,0)	$(\sigma^2_{u0}, \sigma^2_{v1}) = (0.2, 0.2)$	$(\beta_{00}, \beta_{01}) = (\beta_{10}, \beta_{11}) = (0.5, 0.5)^\top$
(u,0)	$(\sigma^2_{u0}, \sigma^2_{u1}) = (0.2, 0.8)$	$(\beta_{00}, \beta_{01}) = (\beta_{10}, \beta_{11}) = (0.5, 0.5)^\top$
(u,b)	$(\sigma^2_{u0}, \sigma^2_{u1}) = (0.2, 0.8)$	$(\beta_{00}, \beta_{01}) = (0.5, 0.6)^\top, (\beta_{10}, \beta_{11}) = (0, 1.1)^\top$

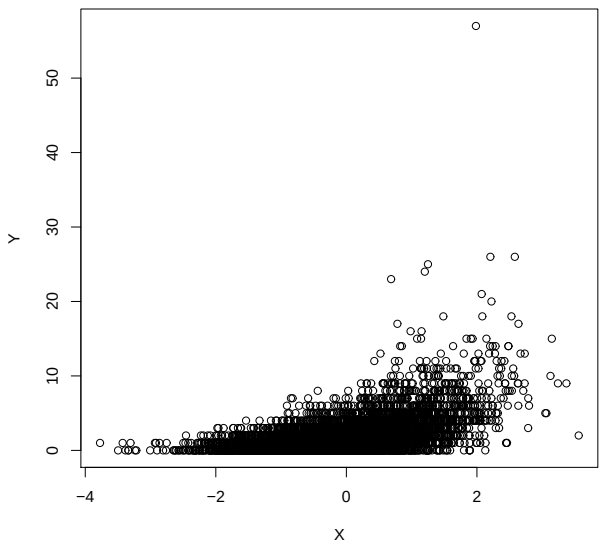
Table 6.2.3: Sources of contamination for Poisson



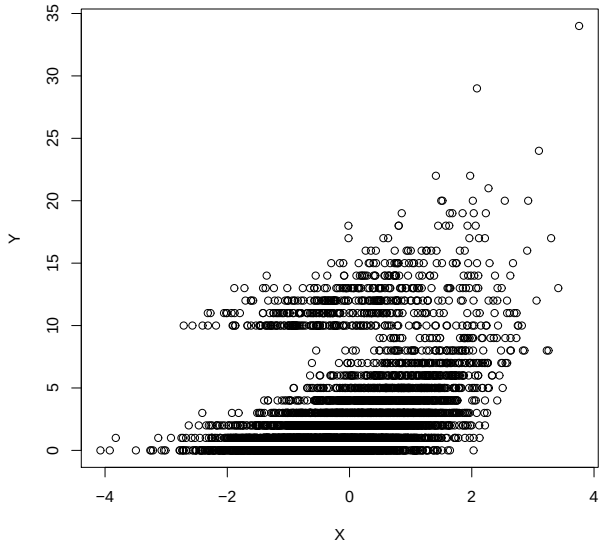
(a) Population 1



(b) Population 2



(c) Population 1



(d) Population 3

Figure 6.2.2: Representation of the four populations

Scenarios	Outliers rate	$\hat{\theta}^{M-Quantile}$	$\hat{\theta}^{REPP}$	$\hat{\theta}^{Syn}$	$\hat{\theta}^{Direct}$
$(0,0)$	0	6.3(166)	2.4(100)	16.2(394)	0.0(240)
$(u,0)$	0.1	4.7(134)	1.9(94)	17.2(292)	0.0(209)
	0.05	6.2(151)	2.3(95)	18.1(365)	0.0(223)
	0.01	8.1(183)	2.8(99)	19.9(465)	0.0(241)
(u,b)	0.1	1.7(78)	1.4(83)	18.1(136)	0.1(204)
	0.05	4.3(92)	1.9(83)	18.3(196)	0.2(208)
	0.01	7.5(152)	2.6(92)	19.6(380)	0.0(229)
Outliers Measurement error	0.1	$-14.3(84)$	$-2.6(90)$	9.5(114)	$-0.1(149)$
	0.05	$-6.1(83)$	$-1.6(85)$	13.2(176)	0.0(165)
	0.01	4.6(140)	1.4(90)	18.3(360)	0.0(212)

Table 6.2.4: Median of the Relative Bias and Relative Efficiency (in the brackets) over all areas

From Table 6.2.4, we first notice that the median bias of the direct estimator is very small, but since the size of the area is small its median mean square error is larger than the EPP median mean square error with values of relative efficiency ranging from 149% to 212%. We note also that the synthetic predictor is also biased and inefficient under all scenario with a large median bias ranging from 9.49% to 19.9% and median relative efficiency values ranging from 114% to 465%. We now focus on the two robust predictors and comment their median bias and relative efficiency.

In absence of outlier (under the scenario $(0,0)$), the robust predictor based on the conditional bias was nearly as efficient as the EPP, whereas the M-quantile is slightly less efficient than the EPP . In this case the value of the tuning constant was large, and the proposed predictor $\hat{\theta}^{REPP}$ was close to the EPP. Under the scenario $(u,0)$, there was a small gain of efficiency for the robust predictor based on conditional bias with values ranging from 94 to 99 whereas the M-quantile predictor was less efficient than the EPP with values ranging from 134% to 183%. Under the scenario (u,b) and in the case of outliers measurement error, we note significant improvements for $\hat{\theta}^{REPP}$ with values of the median efficiency ranging from 83% to 92% for any outlier rate. We noted also some significant improvements for

the M-quantile predictor for outlier rate equal to 5% or 10% with corresponding median values of efficiency equal to 92% and 78%.

6.3 Final remarks

In this chapter, we proposed an extension of Beaumont et al. (2013) to a model-based approach with the use of GLM and showed empirically that the robust predictor performed very well in terms of MSE. We also focused on small area prediction for binary and count data. We proposed an extension of the results of Dongmo Jiongo et al. (2013) involving the conditional bias to binary and count data with an adaptative tuning constant. The estimation of the MSE of the proposed robust predictor is an area of current research.

Appendix

Details of the approximation of the conditional bias of the EBP

We show that an approximation of the conditonal bias of the EBP attached to unit i is given by

$$B_i^{EBP} = \begin{cases} \sum_{j\in U\setminus s} h_j^{'}(\eta_j)\boldsymbol{x}_j^{\top} E_m \left[\left\{\sum_{k\in s} \boldsymbol{H}\left(y_k,\boldsymbol{\beta}\right)\right\}^{-1} \sum_{k\in s} \boldsymbol{t}(y_k,\boldsymbol{\beta}) \mid s, Y_i = y_i\right] & \text{if } i \in s \\ \sum_{j\in U\setminus s} h_j^{'}(\eta_j)\boldsymbol{x}_j^{\top} E_m \left[\left\{\sum_{j\in s} \boldsymbol{H}\left(y_j,\boldsymbol{\beta}\right)\right\}^{-1} \sum_{k\in s} \boldsymbol{t}\left(y_k,\boldsymbol{\beta}\right)\right] - \left\{y_i - h\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)\right\} & \text{if } i \in U\setminus s \end{cases}$$

Following Fuller (2011, page 65), assuming that $\boldsymbol{t}(y_k, .)$ and $\boldsymbol{H}\left(y_k, .\right)$ are continuous on a close set $\mathcal{B}$ containing $\boldsymbol{\beta}$ as interior point and that the Hessian matrix verifies for all $\boldsymbol{\beta}^0$ in $\mathcal{B}$

$$\frac{1}{n}\sum_{k\in s} \boldsymbol{H}\left(y_k,\boldsymbol{\beta}^0\right) = \frac{1}{N}\sum_{k\in U} \boldsymbol{H}\left(y_k,\boldsymbol{\beta}^0\right) + O_p\left(\frac{1}{n^{1/2}}\right)$$

with $\lim_{N\to\infty} \frac{1}{N}\sum_{k\in U} \boldsymbol{H}\left(y_k,\boldsymbol{\beta}^0\right) = H(\boldsymbol{\beta}^0)$ where $H(\boldsymbol{\beta}^0)$ is non singular and assuming that $\|\boldsymbol{t}(y_k, .)\| < K(y_k)$ for some $K(.)$ with finite fourth moment for all y_k and assuming that $\boldsymbol{\beta}$ converges in probability to $\boldsymbol{\beta}$, we have

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \;=\; \left\{\sum_{j\in s} \boldsymbol{H}\left(y_j,\boldsymbol{\beta}\right)\right\}^{-1} \sum_{k\in s} \boldsymbol{t}(y_k,\boldsymbol{\beta}) + o_p\left(\frac{1}{n^{1/2}}\right). \qquad \text{(A.1)}$$

Using a first-order Taylor expansion of the function $h(.)$ assumed differentiable with continuous derivative, we have

$$h\left(\boldsymbol{x}_i^{\top}\hat{\boldsymbol{\beta}}\right) \;=\; h\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right) + h_i^{'}(\eta_i)\boldsymbol{x}_i^{\top}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + o_p\left(\frac{1}{n^{1/2}}\right). \qquad \text{(A.2)}$$

Then combining (A.1) and (A.2), this leads to

$$h\left(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\right) - h\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right) = h_i^{'}(\eta_i)\boldsymbol{x}_i^\top \left\{\sum_{j\in s} \boldsymbol{H}\left(y_j, \boldsymbol{\beta}\right)\right\}^{-1} \sum_{k\in s} \boldsymbol{t}\left(y_k, \boldsymbol{\beta}\right) + o_p\left(\frac{1}{n^{1/2}}\right). \quad \text{(A.3)}$$

The conditional bias of a non-sample unit can be expressed as

$$\begin{aligned} E_m\left(\hat{\theta}^{EBP} - \theta \,|\, s, Y_i = y_i\right) &= E_m\left\{\sum_{j\in s} Y_j + \sum_{j\in U\setminus s} h\left(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}\right) - \sum_{j\in U} Y_j \,|\, s, Y_i = y_i\right\} \\ &= E_m\left[\sum_{j\in U\setminus s}\left\{h\left(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}\right) - h\left(\boldsymbol{x}_j^\top \boldsymbol{\beta}\right)\right\} + \sum_{j\in U\setminus s}\left\{h\left(\boldsymbol{x}_j^\top \boldsymbol{\beta}\right) - Y_j\right\} \,|\, s, Y_i = y_i\right] \end{aligned}$$

Using the expression (A.3) and ignoring the higher order terms, we have

$$\begin{aligned} E_m\left(\hat{\theta}^{EBP} - \theta \,|\, s, Y_i = y_i\right) &= E_m\left[\sum_{j\in U\setminus s} h_j^{'}(\eta_j)\boldsymbol{x}_j^\top \left\{\sum_{j\in s} \boldsymbol{H}\left(y_j, \boldsymbol{\beta}\right)\right\}^{-1} \sum_{k\in s} \boldsymbol{t}\left(y_k, \boldsymbol{\beta}\right)\right. \\ &\quad \left. + \sum_{j\in U\setminus s}\left\{h\left(\boldsymbol{x}_j^\top \boldsymbol{\beta}\right) - y_j\right\} \,|\, s, Y_i = y_i\right] \\ &= \sum_{j\in U\setminus s} h_j^{'}(\eta_j)\boldsymbol{x}_j^\top E_m\left[\left\{\sum_{j\in s} \boldsymbol{H}\left(y_j, \boldsymbol{\beta}\right)\right\}^{-1} \sum_{k\in s} \boldsymbol{t}\left(y_k, \boldsymbol{\beta}\right)\right] - \left\{y_i - h\left(\boldsymbol{x}_j^\top \boldsymbol{\beta}\right)\right\}. \end{aligned} \quad \text{(A.4)}$$

We now develop the conditional bias of a sampled unit i.

$$\begin{aligned} B_i^{EBP}(I_i=1) &= E_m\left(\hat{\theta}^{EBP}-\theta|s,Y_i=y_i\right) \\ &= E_m\left\{\sum_{j\in s}Y_j+\sum_{j\in U\setminus s}h\left(\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}\right)-\sum_{j\in U}Y_j\,|\,s,Y_i=y_i\right\} \\ &= E_m\left[\sum_{j\in U\setminus s}\left\{h\left(\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}\right)-h\left(\boldsymbol{x}_j^\top\boldsymbol{\beta}\right)\right\}+\sum_{j\in U\setminus s}\left\{h\left(\boldsymbol{x}_j^\top\boldsymbol{\beta}\right)-Y_j\right\}\,|\,s,Y_i=y_i\right] \end{aligned}$$

Since $E_m\left[\sum_{j\in U\setminus s}\left\{h\left(\boldsymbol{x}_j^\top\boldsymbol{\beta}\right)-Y_j\right\}\,|\,s,Y_i=y_i\right]=0$ and using the first order Taylor approximation given by (A.3) we have the following approximation of the conditional bias of the EBP attached to unit i

$$\begin{aligned} B_i^{EBP}(I_i=1) &= E_m\left[\sum_{j\in U\setminus s}h_j^{'}(\eta_j)\boldsymbol{x}_j^\top\left\{\sum_{j\in S}\boldsymbol{H}\left(y_j,\boldsymbol{\beta}\right)\right\}^{-1}\sum_{k\in s}\boldsymbol{t}(y_k,\boldsymbol{\beta})\,|\,s,Y_i=y_i\right] \\ &= \sum_{j\in U\setminus s}h_j^{'}(\eta_j)\boldsymbol{x}_j^\top E_m\left[\left\{\sum_{j\in S}\boldsymbol{H}\left(y_j,\boldsymbol{\beta}\right)\right\}^{-1}\sum_{k\in s}\boldsymbol{t}(y_k,\boldsymbol{\beta})\,|\,s,Y_i=y_i\right]. \end{aligned} \tag{A.5}$$

Proof of Proposition 6.1 : we show that the decomposition of the prediction error $\hat{\theta}^{BLUP}-\theta$ is approximately equal to $\sum_{i\in U\setminus s} B_i^{EBP}(I_i=0)+\sum_{i\in s} B_i^{EBP}(I_i=1)$.

By replacing $B_i^{EBP}(I_i=0)$ and $B_i^{EBP}(I_i=1)$ by their respective expression (A.4) and (A.5), we have

$$
\begin{aligned}
& \sum_{i\in U\setminus s} B_i^{EBP}(I_i=0)+\sum_{i\in s} B_i^{EBP}(I_i=1) \\
= & -\sum_{i\in U\setminus s}\left\{y_i-h\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)\right\}+\sum_{i\in s}\sum_{j\in U\setminus s} h_j^{'}(\eta_j)\boldsymbol{x}_j^\top\left\{\sum_{k\in s}\boldsymbol{H}(\boldsymbol{\beta})\right\}^{-1}\boldsymbol{x}_i w_i g_i^{'}(\mu_i)\left\{y_i-h\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)\right\} \\
= & -\sum_{i\in U\setminus s}\left[y_i-\left\{h\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)-h\left(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}\right)\right\}-h\left(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}\right)\right] \\
+ & \sum_{j\in U\setminus s} h_j^{'}(\eta_j)\boldsymbol{x}_j^\top\left\{\sum_{k\in s}\boldsymbol{H}(\boldsymbol{\beta})\right\}^{-1}\sum_{i\in s}\boldsymbol{x}_i w_i g_i^{'}(\mu_i)\left\{y_i-h\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)\right\} \\
= & -\sum_{j\in U\setminus s} Y_j+\sum_{j\in U\setminus s} h\left(\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}\right) \\
- & \sum_{i\in U\setminus s} h_i^{'}(\eta_i)\boldsymbol{x}_i^\top\left\{\sum_{j\in s}\boldsymbol{H}(y_j,\boldsymbol{\beta})\right\}^{-1}\sum_{k\in s}\boldsymbol{x}_k w_k g_k^{'}(\mu_k)\left\{y_k-h\left(\boldsymbol{x}_k^\top\boldsymbol{\beta}\right)\right\} \\
+ & \sum_{j\in U\setminus s} h_j^{'}(\eta_j)\boldsymbol{x}_j^\top\left\{\sum_{k\in s}\boldsymbol{H}(\boldsymbol{\beta})\right\}^{-1}\sum_{i\in s}\boldsymbol{x}_i w_i g_i^{'}(\mu_i)\left\{y_i-h\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)\right\}+o_p\left(\frac{1}{n^{1/2}}\right) \\
= & -\sum_{j\in U\setminus s} Y_j+\sum_{j\in U\setminus s} h\left(\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}\right)+o_p\left(\frac{1}{n^{1/2}}\right) \\
= & \hat{\theta}^{BLUP}-\theta+o_p\left(\frac{1}{n^{1/2}}\right)
\end{aligned}
$$

□

Details of the approximation of the conditional bias of the EPP

We show that the conditional bias of the EPP attached to the unit j in domain h can be express as

$$B_{dhj}\left(y_{hj},u_h;\boldsymbol{\beta}\right)=\begin{cases}N_d^{-1}\left(\sum_{h=1}^{D}\sum_{j\in s_h}w_{dhj}u_h-\sum_{j\in U_d\backslash s_d}h^{'}\left(\eta_{dj}\right)u_d+w_{ddj}e_{dj}\right) & j\in s_d\\ N_d^{-1}\left(\sum_{h=1}^{D}\sum_{j\in s_h}w_{dhj}u_h-\sum_{j\in U_d\backslash s_d}h^{'}\left(\eta_{dj}\right)u_d+w_{dhj}e_{hj}\right) & j\in s_h, h\neq d\\ N_d^{-1}\left(\sum_{h=1}^{D}\sum_{j\in S_h}w_{dhj}u_h-\sum_{j\in U_d\backslash s_d}h^{'}\left(\eta_{dj}\right)u_d-\frac{\partial h}{\partial\eta}\left(\eta_{dj}\right)e_{dj}\right) & j\in U_d\backslash s_d\\ N_d^{-1}\left(\sum_{h=1}^{D}\sum_{j\in s_h}w_{dhj}u_h-\sum_{j\in U_d\backslash s_d}h^{'}\left(\eta_{dj}\right)u_d\right) & j\in U_h\backslash s_h, h\neq d,\end{cases}$$

First the EPP can be decomposed as

$$\begin{aligned}\hat{\theta}_d^{EPP} &= \tfrac{1}{N_d}\left\{\sum_{j\in s_d}y_{dj}+\sum_{j\in U_d\backslash s_d}h\left(\boldsymbol{x}_{dj}^{\top}\hat{\boldsymbol{\beta}}+\hat{u}_d\right)\right\}\\ &= \tfrac{1}{N_d}\left[\sum_{j\in s_d}y_{dj}+\sum_{j\in U_d\backslash s_d}\left\{h\left(\boldsymbol{x}_{dj}^{\top}\hat{\boldsymbol{\beta}}+\hat{u}_d\right)-h\left(\boldsymbol{x}_{dj}^{\top}\boldsymbol{\beta}+u_d\right)\right\}+\sum_{j\in U_d\backslash s_d}h\left(\boldsymbol{x}_{dj}^{\top}\boldsymbol{\beta}+u_d\right)\right]\end{aligned}$$

Then using a first order Taylor approximation, we have

$$\hat{\theta}_d^{EPP}\simeq\frac{1}{N_d}\left\{\sum_{j\in s_d}y_{dj}+\sum_{j\in U_d\backslash s_d}h\left(\boldsymbol{x}_{dj}^{\top}\boldsymbol{\beta}+u_d\right)+A\right\}$$

where

$$
\begin{aligned}
A &= \sum_{j\in U_s\setminus s_d} \frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)\left(\boldsymbol{x}_{dj}^{\top}\hat{\boldsymbol{\beta}} + \hat{u}_d - \boldsymbol{x}_{dj}^{\top}\boldsymbol{\beta} - u_d\right) \\
&= \sum_{j\in U_d\setminus s_d} \frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)\left[\boldsymbol{x}_{dj}^{\top}\left\{\left(\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\boldsymbol{X}_h\right)^{-1}\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\boldsymbol{\xi}_h - \boldsymbol{\beta}\right\}\right] \\
&+ \sum_{j\in U_d\setminus s_d} \frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)\left[\sigma_u^2\mathbf{1}_{n_d}^{\top}\boldsymbol{V}_d^{-1}\left\{\boldsymbol{\xi}_d - \boldsymbol{X}_d\left(\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\boldsymbol{X}_h\right)^{-1}\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\boldsymbol{\xi}_h\right\} - u_d\right] \\
&= -\sum_{j\in U_d\setminus s_d} \frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)u_d + \sum_{j\in U_d\setminus s_d} \frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)\left[\boldsymbol{x}_{dj}^{\top}\left\{\left(\sum_{h=1}^{D}\boldsymbol{X}_h^{T}\boldsymbol{V}_h^{-1}\boldsymbol{X}_h\right)^{-1}\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\boldsymbol{\xi}_h - \boldsymbol{\beta}\right\}\right] \\
&+ \sum_{j\in U_d\setminus s_d} \frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)\left(\sigma_u^2\mathbf{1}_{n_d}^{\top}\boldsymbol{V}_d^{-1}\left[\boldsymbol{\xi}_d - \boldsymbol{X}_d\left(\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\boldsymbol{X}_h\right)^{-1}\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\boldsymbol{\xi}_h\right]\right)
\end{aligned}
$$

and

$$
\boldsymbol{\xi}_h = \boldsymbol{X}_h\boldsymbol{\beta} + \boldsymbol{u}_h + \boldsymbol{e}_h. \tag{A.6}
$$

Using the expression A.6, the term A reduces to

$$
\begin{aligned}
A &= -\sum_{j\in U_d\setminus s_d} \frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)u_d + \sum_{j\in U_d\setminus s_d} \frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)\left[\boldsymbol{x}_{dj}^{\top}\left\{\left(\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\boldsymbol{X}_h\right)^{-1}\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\left(\boldsymbol{u}_h + \boldsymbol{e}_h\right)\right\}\right] \\
&+ \sum_{j\in U_d\setminus s_d} \frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)\left[\sigma_u^2\mathbf{1}_{n_d}^{\top}\boldsymbol{V}_d^{-1}\left\{\boldsymbol{u}_d + \boldsymbol{e}_d - \boldsymbol{X}_d\left(\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\boldsymbol{X}_h\right)^{-1}\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\left(\boldsymbol{u}_h + \boldsymbol{e}_h\right)\right\}\right] \\
&= -\sum_{j\in U_d\setminus s_d} \frac{\partial h}{\partial \eta}\left(\eta_{dj}\right)u_d + B
\end{aligned}
$$

Noting that $\sum_{h=1}^{D}\boldsymbol{X}_h^{\top}\boldsymbol{V}_h^{-1}\left(\boldsymbol{u}_h + \boldsymbol{e}_h\right) = \sum_{h=1}^{D}\sum_{j\in s_h}\boldsymbol{X}_h^{\top}\boldsymbol{C}_h^{(j)}\left(u_h + e_{hj}\right)$ where $C_h^{(j)}$ corresponding to the $j-th$ column of $\boldsymbol{C}_h = \boldsymbol{V}_h^{-1}$, it can be shown that the

second term B can be expressed as :

$$B = \sum_{h=1}^{D} \sum_{j \in s_h} w_{dhj} \left(u_h + e_{hj} \right)$$

where

$$w_{dhj} = \begin{cases} D^{-1} a_d \boldsymbol{X}_h^\top \boldsymbol{C}_h^{(j)} & j \in s_h \\ D^{-1} a_d \boldsymbol{X}_d^\top \boldsymbol{C}_d^{(j)} + \left[\sum_{j' \in U_d \setminus s_d} \frac{\partial h}{\partial \eta} \left(\eta_{dj'} \right) \right] \sigma_u^2 \mathbf{1}_{n_d}^\top \boldsymbol{C}_d^{(j)} & j \in s_d, \end{cases}$$

and

$$a_d = \left\{ \sum_{j \in U_d \setminus s_d} \frac{\partial h}{\partial \eta} \left(\eta_{dj} \right) \left(\boldsymbol{x}_{dj}^\top - \sigma_u^2 \mathbf{1}_{n_d}^\top \boldsymbol{V}_d^{-1} \boldsymbol{X}_d \right) \right\} \left(D^{-1} \sum_{h=1}^{D} \boldsymbol{X}_h^\top \boldsymbol{V}_h^{-1} \boldsymbol{X}_h \right)^{-1}.$$

Then

$$\begin{aligned} \hat{\theta}_d^{EPP} - \theta_d &= \frac{1}{N_d} \left[- \sum_{j \in U_d \setminus s_d} \left\{ y_{dj} - h \left(\boldsymbol{x}_{dj}^\top \boldsymbol{\beta} + u_d \right) + \frac{\partial h}{\partial \eta} \left(\eta_{dj} \right) u_d \right\} + \sum_{h=1}^{D} \sum_{j \in s_d} w_{dhj} \left(u_h + e_{hj} \right) \right] \\ &= \frac{1}{N_d} \left[- \sum_{j \in U_d \setminus s_d} \frac{\partial h}{\partial \eta} \left(\eta_{dj} \right) \frac{\partial g}{\partial \mu} \left(\mu_{dj} \right) \left\{ y_{dj} - h \left(\boldsymbol{x}_{dj}^\top \boldsymbol{\beta} + u_d \right) \right\} + u_d + \sum_{h=1}^{D} \sum_{j \in s_d} w_{dhj} \left(u_h + e_{hj} \right) \right] \\ &= \frac{1}{N_d} \left\{ - \sum_{j \in U_d \setminus s_d} \frac{\partial h}{\partial \eta} \left(\eta_{dj} \right) \left(e_{dj} + u_d \right) + \sum_{h=1}^{D} \sum_{j \in s_h} w_{dhj} \left(u_h + e_{hj} \right) \right\} \end{aligned} \tag{A.7}$$

Applying the definition of the conditional bias, i.e $B_{dhj} \left(y_{hj}, u_h; \boldsymbol{\beta} \right) = E_m \left\{ \hat{\theta}_d^{EPP} - \theta_d | s,\, y_{hj},\, \boldsymbol{u} \right\}$, leads to the result.

References

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555–569.

Cantoni, E., and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96, 1022–1030..

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063–1069.

Chambers, R., Salvati, N. and Tzavidis, N. (2014) *M-quantile regression models for binary data in small area estimation.*

Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 47–69

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255–268.

Clark, R.G. (1995). Winsorization methods in sample surveys. Masters thesis, Department of Statistics, Australian National University.

Dongmo Jiongo, V., Haziza, D. and Duchesne, P. (2013). Controlling the bias of robust small area estimators. *Biometrika*, 100, 843–858.

Fabrizi, E., Salvati, N., Pratesi, M. and Tzavidis, N. (2014). Outlier robust model-assisted small area estimation. *Biometrical Journal*, 56, 157–175.

Fellner, W.H. (1986). Robust estimation of variance components. *Technometrics*, 28, 51–60.

Fuller, W. A. (2011). *Sampling statistics.* Wiley.

Ghosh, M., Natarajan, K., Stroud, T. and Carlin, B. P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273–282.

Ghosh, M., Maiti,T. and Roy, A. (2008). Influence functions and robust Bayes and empirical Bayes small area estimation. *Biometrika*, 93, 255–268.

González-Manteiga, W., Lombardia, M. J., Molina, I., Morales, D. and Santamaria, L. (2007) Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational statistics & data analysis*, 51, 2720–2733.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111, 117–127.

Jiang, J. and Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53, 217–243.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST*, 15, 1–96.

Molina, I., and Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38, 369–385.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M. and Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: conditional bias. *Biometrika*, 86, 923–928.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M., Jimenez-Gamero, M.D. and Muñoz-Pichardo, J. (2002). Influence diagnostics in survey sampling: estimating the conditional bias. *Metrika*, 55, 209–214.

Munoz-Pichardo, J., Munoz-Garcia, J., Moreno-Rebollo, J. and Pino-Mejias, R. (1995). A new approach to influence analysis in linear models. *Sankhyā: The Indian Journal of Statistics, Series A*, 393–409.

Rao, J. N.K. (2003). *Small area estimation*. Wiley.

Saei, A. and Chambers, R. (2003). Small area estimation under linear and generalized linear mixed models with time and area effects. *S3RI Methodology Working Papers M03/15*, Southampton Statistical Sciences Research Institute.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4), 719–727.

Sinha, S. K. and Rao, J.N.K. (2009) Robust small area estimation. *Canadian Journal of Statistics*, 37, 381-399 .

Stahel, W.A. and Welsh, A. (1997). Approaches to robust estimation in the simplest variance components model. *Journal of Statistical Planning and Inference*, 57, 295–319.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010) Robust estimation of small-area means and quantiles. *Australian & New Zealand Journal of Statistics*, 52, 167–186 .

Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E. and Chambers, R. (2014). Robust small area prediction for counts. *Statistical methods in medical research.*

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000*). Finite population sampling and inference: a prediction approach.* Wiley.

Chapitre 7

Conclusion et perspectives

Au cours de cette thèse, nous avons montré que les unités influentes sont très fréquentes quelque soit le type d'approche considéré. Si on utilise des techniques d'estimation basées sur le plan de sondage, des unités influentes peuvent apparaitre sous la forme de "Stratum Jumpers" si la base de sondage dont on dispose est obsolète. Les unités influentes peuvent également se manifester lorsque la variable d'intérêt n'est peu ou pas liée au variable de stratification, et par conséquent très peu liée au poids d'estimation, produisant alors des estimations très instables. Dans le cas d'une approche modèle, une mauvaise spécification du modèle peut également conduire à la présence de valeurs influentes.

Dans le cadre d'une approche sous le plan, les estimateurs winsorisés sont souvent privilégié en pratique. Même si le choix du seuil intervenant dans ces estimateurs est crucial, la façon dont il est déterminé reste très empirique, mis à part pour quelques plans de sondages simples ou sous des hypothèses de modélisation simplificatrices. C'est pourquoi, nous avons développé dans le premier chapitre de cette thèse, une méthode permettant de déterminer de façon adaptative une constante correspondant au seuil des estimateurs winsorisés pour tout type de plan de sondage. L'estimation pour des domaines est un problème important en

pratique, ainsi nous avons développé une méthode combinant des techniques d'estimation robuste et de calage afin de proposer des estimations robustes au niveau des domaines et au niveau de la population tout en vérifiant la relation de cohérence qui les lient.

Dans le chapitre 2, nous avons étendu les résultats de Beaumont et al. (2013) développé dans le cas d'un plan de sondage à une phase à un plan de sondage à deux phases. Les plans à deux phases sont souvent utilisés en pratique pour des sondages utilisant une base de sondage très pauvre en information auxiliaire. Dans ce cas, il est courant de sélectionner un premier échantillon de taille importante pour récupérer de l'information auxiliaire basique liée aux variables d'intérêt de l'enquête. En utilisant les variables observées sur le premier échantillon, on peut alors développer une stratégie de tirage efficace pour sélectionner au sein du premier échantillon un sous échantillon sur lequel on va récupérer les variables d'intérêt de l'enquête. La théorie des plans à deux phases est très intéressante dans la mesure où l'ensemble des répondants est souvent vu comme la réalisation d'une deuxième phase d'échantillonnage. Nous avons proposé dans ce contexte une version robuste de l'estimateur par double dilatation, puis nous avons élaboré une version robuste de l'estimateur ajusté pour la non-réponse construit à partir d'une modélisation par un modèle logistique des probabilité de réponses. Nous avons également traité en détail le cas d'un estimateur utilisant des classes de repondération, qui est l'estimateur le plus souvent utilisé en pratique. Ce problème très intéressant n'avait à notre connaissance pas été traité dans la littérature.

Dans le chapitre 3, nous nous sommes intéressés au problème de l'estimation de la moyenne dans le cas d'une population infinie asymétrique. Dans ce contexte, la moyenne empirique est un estimateur simple de la moyenne de la population. Elle est optimale pour certaines lois comme la loi normale ou la loi exponentielle, mais très instable pour des lois asymétriques comme les lois de Pareto, Weibull et Lognormal. Si la loi des observations était connue, l'estimateur du maximum de vraisemblance serait asymptotiquement optimal pour la moyenne de la population.

Cependant les estimateurs du maximum de vraisemblance sont très sensibles à une mauvaise spécification du modèle. Par exemple, dans le cas de la loi Lognormal, Myers et Pepin (1990) ont montré empiriquement que l'estimation du maximum de vraisemblance peut avoir une erreur quadratique moyenne plus élevée que la moyenne empirique en présence d'une légère mauvaise spécification du modèle. C'est pourquoi, nous avons développé une version non paramétrique robuste de la moyenne empirique basée sur le biais conditionnel. Nous avons donné les propriétés statistiques basiques de cet estimateur et nous avons obtenu une approximation asymptotique de l'erreur quadratique moyenne de l'estimateur robuste suivant le domaine d'attraction du maximum de la distribution ayant généré les observations. Nous avons également proposé un estimateur de l'erreur quadratique moyenne basé sur les statistiques d'ordre de l'échantillon. Dans ce chapitre, nous avons essentiellement développé le cas de la moyenne de la population en construisant une version robuste de la moyenne empirique. L'approche unifiée basée sur le biais conditionnel permet de construire une version robuste d'autres estimateurs comme ceux du maximum de vraisemblance, si on est capable de calculer explicitement ou empiriquement le biais conditionnel de ceux-ci. Enfin, il est possible de construire des versions robustes d'estimateurs de paramètres plus complexes, tels que le coefficient de régression, par exemple.

L'estimation sur petits domaines a connu un essor important au cours de cette dernière décennie. Les tailles d'échantillon observées dans les domaines étant insuffisantes pour produire une estimation directe fiable, donc on utilise des modèles mixtes permettant de modéliser la spécificité de chaque domaine. Il est nécessaire dans ce contexte de construire des estimateurs robustes à une mauvaise spécification du modèle. Dans le chapitre 4, nous avons proposé une version robuste du prédicteur optimal dans le cas d'un modèle linéaire généralisé, puis nous avons étendu les travaux de Dongmo Jiongo et al. (2013), développés dans le cas d'une estimation sur petits domaines à partir d'un modèle linéaire mixte à des modèles linéaires mixtes généralisés. Dans ce dernier chapitre, nous avons supposé

que l'information auxiliaire était disponible au niveau des individus présents dans les domaines. En l'absence d'une information auxiliaire aussi riche, on a recours, par exemple, à des modèles de type Fay-Herriot. Une première extension naturelle des travaux de ce dernier chapitre serait de proposer des versions robustes d'estimateurs issus d'une modélisation au niveau des domaines.

L'objectif de cette thèse était de proposer une méthode d'estimation robuste qui s'adapte aux différentes approches utilisées en Théorie des Sondages, mais aussi à l'estimation en population finie. L'avantage de la méthode d'estimation robuste proposée, basée sur le biais conditionnel, est son adaptabilité au plan de sondage mis en œuvre, au paramètre d'intérêt considéré et à l'estimateur utilisé. Au cours des différents chapitres, nous avons montré que la version robuste construit à partir du biais conditionnel était tout aussi efficace que l'estimateur non robuste en l'absence d'unités influentes et que l'estimateur robuste était plus efficace en terme d'erreur quadratique moyenne en présence d'unités influentes. L'estimation de l'erreur quadratique moyenne dans le cas des sondages de l'estimateur robuste demeure un problème difficile et fera l'objet de recherches ultérieures. Nous avons par ailleurs des extensions des estimateurs robustes basés sur le biais conditionnel un contexte où l'on souhaite faire des prédictions, par exemple sur des séries temporelles d'actifs. Dans ce contexte, l'utilisation d'estimateurs M engendrera un biais important dans les prédictions, surtout si l'objectif du statisticien est de modéliser le comportement de l'ensemble des observations et pas seulement la majorité des observations.